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

Zhenyu Liao
School of Electronic Information & Communication
Huazhong University of Science & Technology, Wuhan, China
zhenyu_liao@hust.edu.cn

Michael W. Mahoney
ICSI, LBNL, and Department of Statistics
University of California, Berkeley, USA
mmahoney@stat.berkeley.edu

March 30, 2025

Contents

Part I

Mathematical Background

In this part, we provide a brief review of the mathematical background that will be used in the remainder of this monograph. This part assumes basic knowledge of the readers, and it aims to present well-known or relatively well-known results in a way that will be particularly useful for our subsequent discussions.

High-dimensional Equivalent

Definition 1.1 (High-dimensional Equivalent). *Let $\phi(\mathbf{X})$ be a nonlinear model of a random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, and let $f(\phi(\mathbf{X}))$ be a 1-Lipschitz scalar observation map with entrywise $\phi\colon \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times n}$ and observation map $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}$. We say that $\bar{\mathbf{X}}_\phi$ (which can be deterministic or random) is a* High-dimensional Equivalent *of $\phi(\mathbf{X})$ with respect to $f(\cdot)$ if, with probability at least $1-\delta(p,n)$ we have that*

$$|f(\phi(\mathbf{X})) - f(\mathbf{X}_\phi)| \leq \varepsilon(n,p), \tag{1.1}$$

for some non-negative functions $\varepsilon(n,p)$ and $\delta(n,p)$ that decrease to zero as $n,p \to \infty$. We denote the relation in (1.1) as

$$\phi(\mathbf{X}) \overset{f}{\leftrightarrow} \mathbf{X}_\phi. \tag{1.2}$$

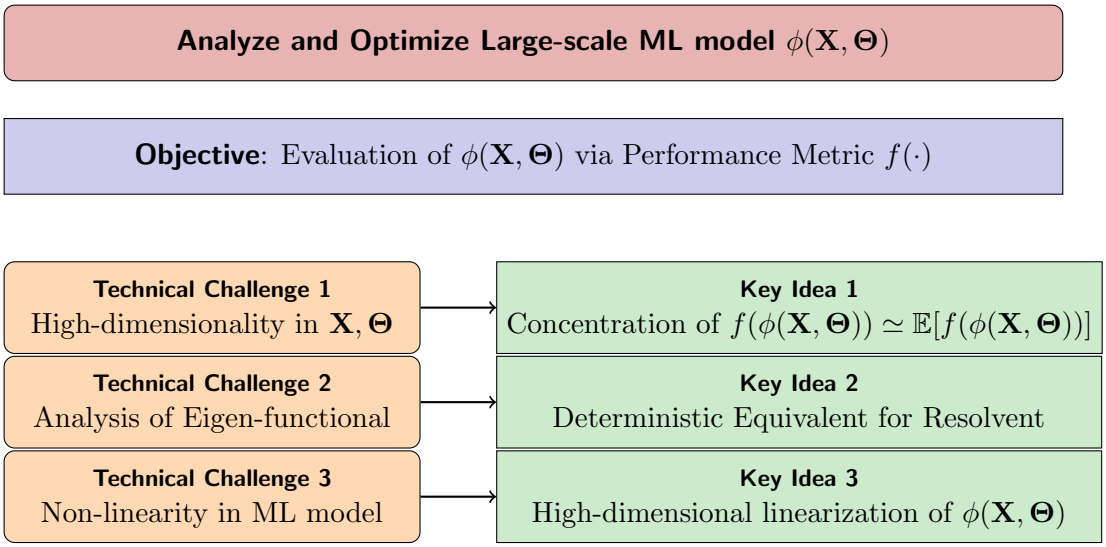


Figure 1.1: Flow diagram of the proposed RMT-based analysis framework for large-scale ML models.

Chapter 1

Basic probability: Random scalars and random vectors

In this chapter, we briefly review some basic probability results to be used throughout the monograph. In Chapter 1.1, we recall the definition of moment and tail of a scalar random variable, as well as the definition of the class of sub-gaussian and sub-exponential distributions. In Chapter 1.2, we consider the sample mean of a collection of independent random variables, and we review its asymptotic characterization via the law of large numbers (LLN) and the central limit theorem (CLT). In Chapter 1.3, we view the sample mean as a *linear scalar observation* $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ of a large random vector $\mathbf{x} \in \mathbb{R}^n$, and we establish non-asymptotic concentration results on $f(\mathbf{x})$. In Chapter 1.4, we extend the concentration results on linear scalar observations (of $\mathbf{x} \in \mathbb{R}^n$) to Lipschitz and even certain non-Lipschitz observations. Finally, in Chapter 1.5, we give a preview of how similar concentration behaviors will extend to large-dimensional random matrices.

1.1 Scalar random variables: moments and tails

Let us start with a scalar random variable. Given a scalar random variable $x \in \mathbb{R}$, one can characterize its behavior via its distribution/law. Equivalently, one can characterize its behavior via its successive moments (when they are well defined) or its moment generating function (MGF). In particular, the MGF, and/or the successive moments of a random variable x, as well as whether or not it satisfies the sub-gaussian or the sub-exponential property, provide different ways to characterize (the properties of) the law/distribution of x.

The definition of these concepts, as well as their connections, are given as follows.

Definition 1.1 (**Moments and moment generating function, MGF**). *For a scalar random variable x defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote*

1. $\mathbb{E}[x]$ *the* expectation *of x;*
2. $\mathrm{Var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$ *the* variance *of x;*
3. *for $p > 0$, $\mathbb{E}[x^p]$ the p^{th}* moment *of x;*
4. *for $p > 0$, $\mathbb{E}[|x|^p]$ the p^{th}* absolute moment *of x; and*
5. *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.*

The p^{th} (absolute) moment of a scalar random variable x can be written as an integral of the *tail* of that random variable, as follows. The tail is of interest since it provides a characterization of the probability that the random variable x differs from a deterministic value (e.g., its expectation, or zero in the case of Lemma 1.2 below) by more than a certain amount $t > 0$.

Lemma 1.2 (**Moments versus tails**). *For a scalar random variable x and fixed $p > 0$, we have*

$$\mathbb{E}[|x|^p] = \int_0^\infty pt^{p-1}\mathbb{P}\left(|x| \geq t\right)\, dt, \tag{1.1}$$

as long as the right-hand side term is finite; and

$$\mathbb{P}\left(|x| \geq t\right) \leq \exp(-\lambda t) M_{|x|}(\lambda), \quad t, \lambda > 0, \tag{1.2}$$

with $M_{|x|}(\lambda)$ the MGF of $|x|$ that is assumed finite.

Proof of Lemma 1.2. To prove Equation (1.1), note that for $|x|^p > 0$, we have

$$\mathbb{E}[|x|^p] = \mathbb{E}\int_0^\infty 1_{t\leq|x|^p} dt = \mathbb{E}\int_0^\infty 1_{t\leq|x|} pt^{p-1} dt = \int_0^\infty pt^{p-1}\mathbb{E}[1_{t\leq|x|}] dt = \int_0^\infty pt^{p-1}\mathbb{P}(|X| > t) dt.$$

The proof of Equation (1.2) follows, for $\lambda > 0$, directly from the Markov's inequality, as

$$\mathbb{P}\left(|x| \geq t\right) = \mathbb{P}\left(\exp(\lambda|x|) \geq \exp(\lambda t)\right) \leq \frac{\mathbb{E}[\exp(\lambda|x|)]}{\exp(\lambda t)} = \exp(-\lambda t) M_{|x|}(\lambda).$$

□

Equation (1.2) is known as the (exponential) Markov's inequality.

As a consequence of Lemma 1.2, bounding the tail decay $\mathbb{P}\left(|x| \geq t\right)$ is *equivalent* to controlling the (successive) moments of the random variable x. In particular, consider a random variable x such that $\mathbb{E}[x] = \mu$ and $\mathrm{Var}[x] = \sigma^2$. In this case, we have

$$\mathbb{P}(|x - \mu| \geq t\sigma) \leq t^{-2}, \quad t > 0, \tag{1.3}$$

which is known as the Chebyshev's inequality.

Equation (1.3) permits us to state that a random variable lies within some range, with some probability. In particular, if we allow for some failure probability $\delta \in (0, 1)$, then it follows from Equation (1.3) that, with probability at least $1 - \delta$, the random x *must* lie within the range

$$x \in [\mu - \sigma/\sqrt{\delta}, \mu + \sigma/\sqrt{\delta}]. \tag{1.4}$$

Of course, this result may or may not be useful. For example, depending on δ and σ, the size of this interval can be *large* with respect to size of μ (for $\mu \approx \sigma$ and $\delta = 1/2$, say).

In what follows, we will be particularly interested in the family of sub-gaussian and sub-exponential random variables, i.e., those having tails akin to standard Gaussian and exponential random variables, respectively. Here is the definition of the sub-gaussian distribution.

Sub-gaussian distribution

Definition 1.3 (**Sub-gaussian distribution**). *For a standard Gaussian random variable $x \sim \mathcal{N}(0, 1)$, we have that the law of x is given by the Gaussian measure $\mu(dt) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$, so that*

$$\mathbb{P}(x \geq X) = \mu([X, \infty)) = \frac{1}{\sqrt{2\pi}}\int_X^\infty \exp(-t^2/2)\, dt \leq \exp(-X^2/2). \tag{1.5}$$

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*

$$\mathbb{P}\left(|y| \geq t\right) \leq \exp(-t^2/\sigma_{\mathcal{N}}^2), \tag{1.6}$$

for some $\sigma_{\mathcal{N}} > 0$ (known as the sub-gaussian norm *of y) for all $t > 0$.*

A closely related family is the sub-exponential distribution.

Sub-exponential distribution

Definition 1.4 (**Sub-exponential distribution**). *For an exponential random variable* $x \sim \text{Exp}(\lambda)$ *of parameter* $\lambda > 0$*, we have that the law of* x *is given, for* $X \geq 0$*, by*

$$\mathbb{P}(x \geq X) = \lambda \int_X^\infty \exp(-\lambda t)\, dt = \exp(-\lambda X), \tag{1.7}$$

and 1 *for* $X < 0$*. We say* y *is a* sub-exponential random variable *if it has a tail that decays* as fast as *exponential random variables, that is*

$$\mathbb{P}(|y| \geq t) \leq \exp(-t/\sigma_{\mathcal{N}}). \tag{1.8}$$

for some $\sigma_{\mathcal{N}} > 0$ *(known as the* sub-exponential norm *of* y*) for all* $t > 0$*.*

Clearly, a sub-exponential random variable is somewhat more heavy-tailed than a sub-gaussian random variable, in the sense that it has more probability mass far out in the tail.

We can compare the sub-gaussian tail in Definition 1.3 with the tail bound in Equation (1.3) (which we recall relies only on the assumption of bounded variance): for a sub-gaussian random variable x of mean $\mu = \mathbb{E}[x]$ and sub-gaussian norm $\sigma_{\mathcal{N}}$, one has that

$$\mathbb{P}\left(|x - \mu| \geq t\sigma_{\mathcal{N}}\right) \leq \exp(-t^2), \tag{1.9}$$

for all $t > 0$. From this, we see that 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).

Remark 1.5 (**Concentration of scalar random variables around their means**). Clearly, Equation (1.9) characterizes a *much* stronger concentration behavior than Equation (1.3). Relatedly, Equation (1.9) can also be used to state that a random variable lies within some range, with some probability: with probability at least $1 - \delta$, one has

$$x \in [\mu - \sigma_{\mathcal{N}}\sqrt{\ln(1/\delta)}, \mu + \sigma_{\mathcal{N}}\sqrt{\ln(1/\delta)}]. \tag{1.10}$$

In this case, Equation (1.10) provides much stronger control on the location of x than Equation (1.4). See Figure 1.1a for an illustration of this "concentration around means" behavior of sub-gaussian random variables.

We note, however, that even under this much stronger control of sub-gaussianity, a "tradeoff" exists in Equation (1.10) (and, of course, also in the weaker form in Equation (1.4)) between the *confidence* and the *range* of the random x: increasing the confidence of the estimate (by taking smaller δ) will lead to fluctuation on a larger interval. Due to this tradeoff, the scale of the width of the interval need *not* be small, compared to the mean; and thus it is a priori *inappropriate* to say that the value of the random x can be well approximated by any *deterministic* value, e.g., its expectation $\mu = \mathbb{E}[x]$.

As a concrete example, taking $\delta = 0.01$ and $\sigma_{\mathcal{N}} = \mu$, it follows from Equation (1.10) that the sub-gaussian random x is within the interval $[-3.6\mu, 4.6\mu]$ with confidence 0.99. This is much stronger than Equation (1.4), but may still not be satisfactory in scenarios that are extremely sensitive to approximation errors.

Remark 1.6 (**Sub-gaussian and sub-exponential random vectors**). The idea in Definition 1.3 and Definition 1.4 extends to random vectors. In particular, we say that

1. a random vector $\mathbf{x} \in \mathbb{R}^n$ is *sub-gaussian* if its one dimensional marginals $\mathbf{x}^\mathsf{T}\mathbf{y}$ are, for all $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\| = 1$, sub-gaussian random variables, that is, $\mathbb{P}(|\mathbf{x}^\mathsf{T}\mathbf{y}| \geq t) \leq \exp(-t^2/C_n^2)$ for all $t \geq 0$ and some $C_n > 0$, where that C_n *may depend* on the dimension n; and

2. a random vector $\mathbf{x} \in \mathbb{R}^n$ is *sub-exponential* if its one dimensional marginals $\mathbf{x}^\mathsf{T}\mathbf{y}$ are, for all $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\| = 1$, sub-exponential random variables, that is, $\mathbb{P}(|\mathbf{x}^\mathsf{T}\mathbf{y}| \geq t) \leq \exp(-t/C_n)$ for all $t \geq 0$ and some (possibly dimension-dependent) $C_n > 0$.

We refer the interested readers to [36, Section 3.4] for discussions on sub-gaussian and sub-exponential random vectors. In what follows, we will *not* be particularly interested in this perspective on random vectors.

1.2 A collection of scalar random variables: from LLN to CLT

Many textbooks on statistics and/or data science start with the (asymptotic) study of sums of independent variables, and in particular with the law of large numbers (LLN) and the central limit theorem (CLT). These will be discussed in this section.

For a collection of independent and identically distributed (i.i.d.)[1] random variables $x_1, \ldots, x_n$ of mean μ and variance σ^2, we have, by independence, that

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n x_i\right] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[x_i] = \frac{\sigma^2}{n}. \tag{1.11}$$

That is, the variance of the sample mean $\frac{1}{n}\sum_{i=1}^n x_i$ is n times smaller than that of each component; and, in particular, it vanishes as $n \to \infty$ (as long as σ^2 does not scale with n). This indicates that for n large, the (random) sample mean *strongly concentrates* around its expectation μ, and thus that it is meaningful to say that the random variable can be approximated by a deterministic quantity. This is in sharp contrast to, e.g., the standard sub-gaussian concentration in Equation (1.10) for which the variance or sub-gaussian norm is *independent* of n.

A formal asymptotic characterization of this concentration behavior is given by the law of large numbers (LNN), given as follows.

Theorem 1.7 (**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*

1. *the sample mean $\frac{1}{n}\sum_{i=1}^n x_i \to \mu$* in probability *as $n \to \infty$, that is, for any $t > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n x_i - \mu\right| \geq t\right) = 0, \tag{1.12}$$

   *known as the **weak law of large numbers (WLLN)**; and*

2. *the sample mean $\frac{1}{n}\sum_{i=1}^n x_i \to \mu$* almost surely *as $n \to \infty$, that is*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n x_i = \mu\right) = 1, \tag{1.13}$$

   *known as the **strong law of large numbers (SLLN)**.*

The strong law of large numbers is technically stronger than the weak law (in characterizing a faster growth rate of the success probability to one as n grows), but the intuition remains the

[1] Note that for the LLN and CLT here, the i.i.d. assumption plays a central role and allows for a large "*degrees of freedom*" in the *large* collection of random variables $x_1, \ldots, x_n$. The i.i.d. assumption can be relaxed to *independent*, however at the cost of some additional control on the higher-order moments, e.g., with Lyapunov's CLT, see [4, Theorem 27.3]. As we shall see later in Chapter 6.1, similar results hold for random matrices.

same. Theorem 1.7 provides asymptotic characterization of what can be called the *first-order* and close-to-deterministic behavior of the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$.

The next result, the well-known central limit theorem (CLT), goes one step further by characterizing the limiting behavior of the *second-order* random fluctuation of the *properly scaled* sample mean around its expectation μ.

Theorem 1.8 (**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*

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu) \geq t\right) \to \frac{1}{\sqrt{2\pi}}\int_{t}^{\infty} e^{-x^2/2}\,dx \tag{1.14}$$

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.

The asymptotic concentration properties of the LLN and CLT can be viewed in a unified way, as we describe in the following remark.

Remark 1.9 (**Concentration of sample mean of a collection of random variables: asymptotic characterization**). The results of the LLN and the CLT in Theorem 1.7 and Theorem 1.8, respectively, can be compactly written as

$$\frac{1}{n}\sum_{i=1}^{n} x_i \simeq \underbrace{\mu}_{O(1)} + \underbrace{\mathcal{N}(0,1)\cdot\sigma/\sqrt{n}}_{O(n^{-1/2})}, \tag{1.15}$$

as $n \to \infty$, for μ, σ both of order $O(1)$. Equation (1.15) makes explicit both the first order and second order behavior of the sample mean of a sequence of i.i.d. random variables $x_1, \ldots, x_n$, with $\mathbb{E}[x_i] = \mu$ and $\mathrm{Var}[x_i] = \sigma^2$, as:

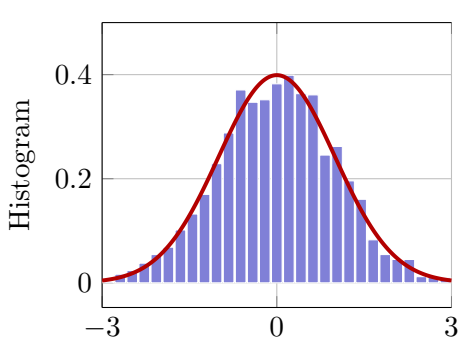
1. in the first order (of magnitude $O(1)$), it has an *asymptotically deterministic* behavior around the expectation μ; and
2. in the second order (of magnitude $O(n^{-1/2})$), it *strongly concentrates* around this deterministic quantity with a *universal* Gaussian fluctuation, *regardless of* the distribution of the component of x_i.

We will see that this behavior occurs well beyond the sum of i.i.d. random variables.

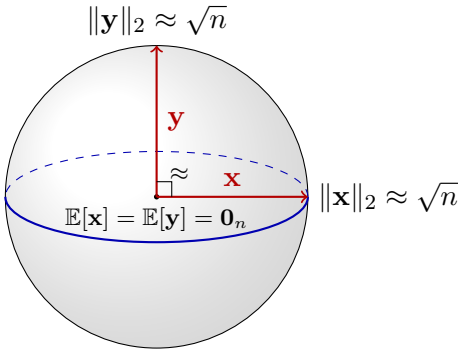
1.3 Concentration of random vectors and their scalar observations

We move on to discuss the concentration properties of random vectors. These have important connections with the results on a collection of random variables in Chapter 1.2; and they will lay the foundations for similar results on random matrices later in Chapter 6. Consider a random vector $\mathbf{x} \in \mathbb{R}^n$ having i.i.d. entries, that is $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$. Without loss of generality, we can choose $\mathbb{E}[x_i] = \mu$, and $\mathrm{Var}[x_i] = \sigma^2$ for $i \in \{1, \ldots, n\}$.

We should first define what we mean by "concentration" for a random vector $\mathbf{x} \in \mathbb{R}^n$. In the following observation, we show that (perhaps rather surprisingly) random vectors do *not* "concentrate" around their means, if we consider the vectors themselves.



(a) "Concentration" around the mean for one-dimensional random vectors



(b) "Non-concentration" around the mean for multi-dimensional random vectors

Figure 1.1: Visualization of the "concentration" (Figure 1.1a) versus "non-concentration" (Figure 1.1) around the mean behavior for one- versus multi-dimensional random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ in XX and Observation 1.10, respectively.

Observation 1.10 (**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

$$\mathbb{E}[\|\mathbf{x} - \mathbf{0}\|_2^2] = \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x}] = \mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}]) = n, \tag{1.16}$$

and further by independence that

$$\mathbb{E}[\|\mathbf{x} - \mathbf{y}\|_2^2] = \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{x} + \mathbf{y}^\mathsf{T}\mathbf{y}] = 2n. \tag{1.17}$$

This means that the origin (which is also the *mean* of $\mathbf{x}$ in this case) is always, in expectation, at the midpoint of two independent draws of random vectors in $\mathbb{R}^n$. The statement easily generalizes to the case of nonzero mean with $\mathbb{E}[\mathbf{x}] \neq \mathbf{0}$, and it allows us to conclude that any random vector $\mathbf{x} \in \mathbb{R}^n$ with n large is *not* close to its mean. More generally, it can be shown that the random vector $\mathbf{x}$ does *not* itself "concentrate" around *any* n-dimensional *deterministic* vector in *any* traditional sense. This large-dimensional counterintuitive "non-concentration" behavior is visualized in Figure 1.1.

In spite of this, from the LLN and CLT in Theorem 1.7 and Theorem 1.8, one expects that some types of "observations" or "measurements" 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, at least as $n \to \infty$. In the following, we can "interpret" the sample mean as a "linear scalar observation" of a vector $\mathbf{x} \in \mathbb{R}^n$.

Remark 1.11 (**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*

$$f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n = \frac{1}{n}\sum_{i=1}^{n} x_i, \;\; \text{or } f(\cdot) = \mathbf{1}_n^\mathsf{T}(\cdot)/n. \tag{1.18}$$

Importantly, while the result of Observation 1.10 shows that a given random vector $\mathbf{x} \in \mathbb{R}^n$ does not itself concentrate/converge in any meaningful sense, Remark 1.11 shows that, *when observed via "linear queries" or scalar observations*, it does concentrate/converge, in this weaker ("scalar observation") sense.

Table 1.1: 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	$f(\mathbf{x}) \to \mu$ LLN in Theorem 1.7	$\frac{\sqrt{n}}{\sigma}(f(\mathbf{x}) - \mu) \to \mathcal{N}(0,1)$ in law CLT in Theorem 1.8
Non-asymptotic under finite variance	$\mathbb{E}[f(\mathbf{x})] = \mu$	$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma/\sqrt{n}\right) \leq t^{-2}$ in Theorem 1.12
Non-asymptotic under sub-gaussianity	$\mathbb{E}[f(\mathbf{x})] = \mu$	$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma_\mathcal{N}/\sqrt{n}\right) \leq \exp(-Ct^2)$ in Theorem 1.13

Let's be clear about what we have done. The quantity $\frac{1}{n}\sum_{i=1}^n x_i$ can be viewed in one of two complementary ways: as the empirical mean of n instantiations of a scalar random variable, providing a meaningful way to quantify how that the empirical mean may concentrate about its population mean; and as a "scalar observation" of a single instantiation of a random vector, which by Observation 1.10 does not concentrate about its mean.

Asymptotic characterization of concentration of linear scalar observations. An *asymptotic* characterization of this concentration behavior for random vectors is given in Equation (1.15) of Remark 1.9, and it is illustrated in Figure 1.1.

Non-asymptotic characterization of concentration of linear scalar observations. We now provide *non-asymptotic* characterizations of the concentration behavior of the linear scalar observation $f(\mathbf{x})$ in Remark 1.11, under two different assumptions on the behavior of the tail of the (entries of the) random vector $\mathbf{x}$. To do so, we consider two cases: that the entries of $\mathbf{x}$

1. are only assumed to have finite variance σ^2 (but nothing is assumed about its tail behavior or higher-order moments); and

2. have sub-gaussian tails with sub-gaussian norm $\sigma_\mathcal{N}$.

The results are summarized in Table 1.1. We now describe them in more detail.

Non-asymptotic analysis of $f(\mathbf{x})$ under finite variance. Let us compute the expectation and variance of the linear scalar observation $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$ of $\mathbf{x} \in \mathbb{R}^n$, for $\mathbf{x}$ having i.i.d. entries with $\mathbb{E}[x_i] = \mu$ and $\mathrm{Var}[x_i] = \sigma^2$:

$$\begin{aligned} \mathbb{E}[f(x)] &= \mathbb{E}[\mathbf{x}^\mathsf{T}\mathbf{1}_n/n] = \mu, \\ \mathrm{Var}[\mathbf{x}^\mathsf{T}\mathbf{1}_n/n] &= \mathbf{1}_n^\mathsf{T}\mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\mathsf{T}]\mathbf{1}_n/n^2 = \sigma^2/n, \end{aligned} \tag{1.19}$$

where we recall that $\mathbb{E}[\mathbf{x}]/\sqrt{n} = \mu\mathbf{1}_n/\sqrt{n}$ and the covariance $\frac{1}{n}\mathrm{Cov}[\mathbf{x}] = \sigma^2\mathbf{I}_n/n$. Note that this nothing but Equation (1.11).

Plugging the results in Equation (1.19) into the Chebyshev's inequality in (1.3), we get the following concentration result for the scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$.

Theorem 1.12 (Concentration of $f(\mathbf{x})$ under finite variance). *For the linear scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ of a random vector $\mathbf{x}$, with $\mathbf{x} \in \mathbb{R}^n$ having i.i.d. entries with $\mathbb{E}[x_i] = \mu$ and $\mathrm{Var}[x_i] = \sigma^2$, we have, for any n and $t > 0$ that*

$$\mathbb{P}\left(\lvert f(\mathbf{x}) - \mu\rvert \geq t\sigma/\sqrt{n}\right) \leq t^{-2}, \tag{1.20}$$

Notably, the linear scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ is *close-to-deterministic* for n large, in the following sense: for some failure probability smaller than δ, it follows from Equation (1.20) that the random x will be within the range of

$$f(\mathbf{x}) \in [\mu - \sigma/\sqrt{n\delta}, \mu + \sigma/\sqrt{n\delta}], \tag{1.21}$$

with probability at least $1-\delta$. Note that here the range is of length $2\sigma/\sqrt{n\delta}$, which, for given σ and δ, can be made *small* for n large. More precisely, consider the case of $\mu = \sigma \neq 0$ and $\delta = 0.01$, having a sufficiently large $n \geq 10^6$ leads to the approximation $f(\mathbf{x}) \in [0.99\mu, 1.01\mu]$ with confidence 0.99. This should be contrasted with Equation (1.4) in which we do *not* observe such large-n concentration for each of the entries of $\mathbf{x}$ to "compensate" the fundamental tradeoff between the confidence and approximation error.

Non-asymptotic analysis of $f(\mathbf{x})$ under sub-gaussianity. Stronger concentration results can obtained under stronger assumptions, e.g., by considering the case of $\mathbf{x}$ having independent sub-gaussian entries. In this case, it follows from the (general) Hoeffding's inequality (see, e.g., [36, Theorem 2.6.2]) that the scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ concentrates within a radius of $1/\sqrt{n}$ from its mean with exponentially high probability, as in the following result.

Theorem 1.13 (Concentration of $f(\mathbf{x})$ under sub-gaussianity). *For the linear scalar observation $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ of a random vector $\mathbf{x}$, with $\mathbf{x} \in \mathbb{R}^n$ having independent sub-gaussian random variables $x_1, \ldots, x_n$ with $\mathbb{E}[x_i] = \mu$ and sub-gaussian norm bounded by $\sigma_\mathcal{N}$, we have, for any n and $t > 0$ that*

$$\mathbb{P}(|f(\mathbf{x}) - \mu| \geq t) \leq \exp(-Cnt^2/\sigma_\mathcal{N}^2), \tag{1.22}$$

or equivalently $\mathbb{P}\left(|f(\mathbf{x}) - \mu| \geq t\sigma_\mathcal{N}/\sqrt{n}\right) \leq \exp(-Ct^2)$ for some constant $C > 0$.

As a consequence of Theorem 1.13, we have that

$$f(\mathbf{x}) \in [\mu - \sqrt{\ln(1/\delta)/C} \cdot \sigma_\mathcal{N}/\sqrt{n}, \mu + \sqrt{\ln(1/\delta)/C} \cdot \sigma_\mathcal{N}/\sqrt{n}] \tag{1.23}$$

with probability at least $1-\delta$. Again, let us compare this expression with Equation (1.10) for a *scalar* sub-gaussian random variable. In Equation (1.23), we do *not* face, for n large, the confidence-range tradeoff observed in Equation (1.10), in the sense that for large n the scale of the width of the interval can be made small, e.g., compared to the mean. As a telling example, consider again the case $\mu = \sigma_\mathcal{N}$ and $\delta = 0.01$, so that by Equation (1.23) we have, for large enough $n \geq 4.6/C \cdot 10^4$, that the approximation $f(\mathbf{x}) \in [0.99\mu, 1.01\mu]$ holds with probability 0.99. Thus, we can confidently say that the value of $f(\mathbf{x})$ can be well-approximated by the deterministic μ.

A few remarks are in order.

Remark 1.14 (Connection to Chernoff bound). In the special case of $\mathbf{x} \in \mathbb{R}^n$ having independent Bernoulli entries (i.e., $\mathbb{P}(x_i = 1) = p_i \in (0,1)$ and $x_i = 0$ otherwise) with $\mu = \frac{1}{n}\sum_{i=1}^n p_i \in (0,1)$, it then follows from the standard Chernoff bound that

$$\mathbb{P}\left(\left|\mathbf{1}_n^\mathsf{T}\mathbf{x}/n - \mu\right| \geq t\right) \leq \exp(-nt^2/(3\mu)), \tag{1.24}$$

for $t \in (0, \mu)$. This agrees with the expression in Equation (1.22) by taking $C = 1/(3\mu)$, since Bernoulli random variables are bounded and thus sub-gaussian.

Here is a summary of the (non-asymptotic) concentration properties of linear scalar observations of large random vectors.

Remark 1.15 (**Concentration of linear scalar observation of large random vectors**)**.** Equation (1.20) and Equation (1.22) of Theorem 1.12 and Theorem 1.13, respectively, show that the random vector $\mathbf{x} \in \mathbb{R}^n$, when "observed" via the linear scalar observation $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$, exhibits the following concentration behavior:

$$f(\mathbf{x}) \simeq \underbrace{\mu}_{O(1)} + \underbrace{X/\sqrt{n}}_{O(n^{-1/2})}, \tag{1.25}$$

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

1. has a tail that decays (at least) as t^{-2}, for finite n and $\mathbf{x}$ having entries of bounded variance (from Equation (1.20));
2. has a sub-gaussian tail (at least) as $\exp(-t^2)$, for finite n and $\mathbf{x}$ having sub-gaussian entries (from Equation (1.22)); and
3. has a precise Gaussian tail *independent* of the law of (the entries of) $\mathbf{x}$, but in the limit of $n \to \infty$ (from the CLT in Theorem 1.8).

To summarize:

1. in the first order (of magnitude $O(1)$), it fluctuates around the *deterministic* quantity μ (that does *not* scale with the dimension n); and
2. in the second order (of magnitude $O(n^{-1/2})$), it exhibits a *strong concentration* around the expectation μ with a fluctuation/deviation (that vanishes as $n^{-1/2}$), the tail behavior of which *depends* on the law of the entries of $\mathbf{x}$.

Remark 1.16 (**Connection between Remark 1.9 and Remark 1.15**)**.** Remark 1.15 takes a similar form to the asymptotic characterization given in Remark 1.9. They both establish *close-to-deterministic* behavior of $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$ with strong concentration, in the sense that the *random fluctuation* is of smaller order than the mean μ. The major differences between the two are the following.

1. Remark 1.9 provides an *asymptotic* characterization of $f(\mathbf{x})$ that holds *only* as $n \to \infty$. However, it holds for more general $\mathbf{x}$, as long as $\mathbf{x}$ has i.i.d. entries of some mean and variance say, in which case the *limiting* fluctuation is precisely Gaussian, as $n \to \infty$.
2. Remark 1.15 provides a *non-asymptotic* characterization of the behavior of $f(\mathbf{x})$ that holds for *any* n. However, this comes at the price of evaluating, on a case by case basis, the precise law of the entries of $\mathbf{x}$.

1.4 Lipschitz, quadratic concentration, and beyond

The discussion around Remark 1.15 was motivated by two facts: Observation 1.10, which noted that random vectors do not concentrate about their mean, in a meaningful manner analogous to how random scalars concentrate about their mean; and that, when working with scalar observations of random vectors, we obtain expressions that are formally equivalent to computing empirical estimates of sums of scalar random variables.

Importantly, the properties described in Remark 1.15 extend beyond the specific *linear* observation, $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$, to many types of (possibly) nonlinear observations. (Clearly, they easily extend to generic linear observations of the form $\mathbf{a}^\mathsf{T}\mathbf{x}$.) Below, we formally define the scalar observation map of (random) vectors.

Table 1.2: 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 $f(\mathbf{x}) = \mathbf{1}_n^\mathsf{T}\mathbf{x}/n$ as in Remark 1.11, and $f(\mathbf{x}) = \mathbf{a}^\mathsf{T}\mathbf{x}$ for $\mathbf{a} \in \mathbb{R}^n$	Table 1.1
Lipschitz	$f(\mathbf{x})$ for a Lipschitz map $f\colon \mathbb{R}^n \to \mathbb{R}$	Theorem 1.19
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 in Theorem 1.22
Nonlinear quadratic form	$f(\mathbf{x}) = \phi(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\phi(\mathbf{Y}^\mathsf{T}\mathbf{x})$ for entry-wise ϕ, $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{Y} \in \mathbb{R}^{p\times n}$	Theorem 1.24

Scalar observation maps

Definition 1.17 (**Scalar observation maps**). *For a (random or not) 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 $f\colon \mathbb{R}^n \to \mathbb{R}$.*

In this section, we will describe several different scalar observation maps (Lipschitz, quadratic functions, and nonlinear quadratic functions), and we will provide characterizations of their concentration behaviors. These results are summarized in Table 1.2.

Lipschitz maps. Consider first a Lipschitz map $f(\mathbf{x})$, for some $f\colon \mathbb{R}^n \to \mathbb{R}$, defined as follow.

Lipschitz function

Definition 1.18 (**Lipschitz function**). *For a function $f\colon \mathbb{R}^n \to \mathbb{R}$, we say f is Lipschitz with Lipschitz constant $K_f > 0$ if*

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \le K_f \|\mathbf{x}_1 - \mathbf{x}_2\|_2, \tag{1.26}$$

holds for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$.

The following result characterizes the concentration behavior of the Lipschitz (scalar) observation of, say, Gaussian random vectors.

Theorem 1.19 (**Concentration of Lipschitz map of Gaussian random vectors, [36, Theorem 5.2.2]**). *For a standard Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and a Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}$ of Lipschitz constant $K_f > 0$, we have, for all $t > 0$ that*

$$\mathbb{P}\left(|f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]| \ge t\right) \le \exp(-Ct^2/K_f^2), \tag{1.27}$$

for some universal constant $C > 0$.

Remark 1.20 (**Concentration of Lipschitz observation of large random vectors**). It follows from Theorem 1.19 that Lipschitz scalar observations $f(\mathbf{x})$ of the random vector $\mathbf{x} \in \mathbb{R}^n$ behave as

$$f(\mathbf{x}) \simeq \mathbb{E}[f(\mathbf{x})] + K_f, \tag{1.28}$$

for n large, where K_f is the Lipschitz constant of f (that is, in general, of order $O(n^{-1/2})$, see Remark 1.21 below). This leads to first- and second-order behaviors akin to those discussed in Remark 1.15 (and Remark 1.9):

1. in the first order, $f(\mathbf{x})$ fluctuates around the deterministic quantity $\mathbb{E}[f(\mathbf{x})]$; and
2. in the second order, it *concentrates* around this deterministic quantity with a fluctuation/deviation that is proportional to K_f and has a sub-gaussian tail.

Remark 1.21 (**Linear observations as Lipschitz observations**). The linear map $\mathbf{1}_n^\mathsf{T}(\cdot)/n$ is, by definition, Lipschitz, with Lipschitz constant $K_{\mathbf{1}_n^\mathsf{T}(\cdot)/n} = n^{-1/2}$. This allows us to deduce the result in Theorem 1.13 from Theorem 1.19 without resorting to the Hoeffding's inequality. More generally, if one has $f(\mathbf{x}) = O(1)$ and

$$|f(\mathbf{y}_1) - f(\mathbf{y}_2)| = O\left(K_f \|\mathbf{y}_1 - \mathbf{y}_2\|_2\right), \tag{1.29}$$

then, for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ (for which we know $\|\mathbf{x}\|_2 = O(\sqrt{n})$), taking $\mathbf{y}_1 = \mathbf{x}$ and $\mathbf{y}_2 = \mathbf{0}$, one deduces that $K_f = O(n^{-1/2})$, so that the second order fluctuation in Equation (1.28) is again of order $O(n^{-1/2})$, as for linear observation in Remark 1.15.

Quadratic form maps. When non-Lipschitz observations of $\mathbf{x}$ are considered (with non-Lipschitz f), one may intuitively expect that the random variable $f(\mathbf{x})$ still concentrates in some way, but "less so," compared to the Lipschitz case. An important special case of this arises when one considers quadratic forms, i.e.,

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \quad \text{for some given} \quad \mathbf{A} \in \mathbb{R}^{n\times n} \quad \text{of} \quad \mathbf{x}, \tag{1.30}$$

from which one can define "quadratic form observations." The following result, known as the *Hanson–Wright inequality*, precisely characterizes the concentration behavior of the *quadratic* (so *non-Lipschitz*) form of $\mathbf{x}$ having independent sub-gaussian entries.

Theorem 1.22 (**Hanson–Wright inequality for quadratic forms, [36, 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*

$$\mathbb{P}\left(\left|\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \operatorname{tr}\mathbf{A}\right| \geq t\right) \leq \exp\left(-\frac{C}{\sigma_\mathcal{N}^2}\min\left(\frac{t^2}{\sigma_\mathcal{N}^2\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|_2}\right)\right), \tag{1.31}$$

for some universal constant $C > 0$.

From Theorem 1.22, we see that, 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. For example, consider $\mathbf{A} = \mathbf{I}_n$ so that $\|\mathbf{A}\|_2 = 1$ and $\|\mathbf{A}\|_F^2 = n$. In this case, it follows from Theorem 1.22 that

$$\mathbb{P}\left(\left|\frac{1}{n}\|\mathbf{x}\|_2^2 - 1\right| \geq \frac{t}{\sqrt{n}}\right) \leq \exp\left(-\frac{C}{\sigma_\mathcal{N}^2}\min\left(\frac{t^2}{\sigma_\mathcal{N}^2}, \sqrt{n}t\right)\right). \tag{1.32}$$

From this, it follows that:

1. close to the mean (that is, equal to one) with $t < \sqrt{n}\sigma_{\mathcal{N}}^2$, the (normalized) squared Euclidean norm $\|\mathbf{x}\|_2^2/n$ *strongly* concentrates around one, with a *sub-gaussian* decay; and
2. away from the mean with $t > \sqrt{n}\sigma_{\mathcal{N}}^2$, the (normalized) squared Euclidean norm $\|\mathbf{x}\|_2^2/n$ still concentrates, but less, with a *sub-exponential* decay.

Remark 1.23 (**Concentration of Euclidean norm of large random vectors**). It follows from Theorem 1.22 that the squared Euclidean norm $\|\mathbf{x}\|_2^2$, as a (non-Lipschitz) quadratic observation of $\mathbf{x} \in \mathbb{R}^n$, behaves as

$$\frac{1}{n}\|\mathbf{x}\|_2^2 \simeq 1 + O(n^{-1/2}), \tag{1.33}$$

for n large. This, again, leads to the first- and second-order behaviors as:

1. in the first order, $\|\mathbf{x}\|_2^2/n$ fluctuates around the deterministic quantity one; and
2. in the second order, it *concentrates* around this deterministic quantity with a fluctuation/deviation that grows with $\sigma_{\mathcal{N}}^2$ and of order $O(n^{-1/2})$ with a *sub-gaussian* tail when close to the deterministic quantity, and with a *sub-exponential* tail (so with a fluctuation with heavier tail and concentrates "less" than the Lipschitz case) when far away.

This should be compared and contrasted with the case of Lipschitz maps in Remark 1.20.

Nonlinear quadratic form maps. More generally, we may be interested in more involved observations of large random vectors than the quadratic forms characterized by the Hanson–Wright inequality in Theorem 1.22. An example is nonlinear quadratic forms of the type

$$\frac{1}{n}\phi(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\phi(\mathbf{Y}^\mathsf{T}\mathbf{x}), \tag{1.34}$$

for Gaussian random $\mathbf{x} \in \mathbb{R}^p$ and deterministic $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{Y} \in \mathbb{R}^{p\times n}$. This is of direct use in the analysis of large and random neural network models in **??**, for which the random vectors $\mathbf{x}$ are (columns of) the network weights applied on deterministic input data $\mathbf{Y}$. A nonlinear *activation function* $\phi\colon \mathbb{R} \to \mathbb{R}$ of a neuron is then applied entry-wise on $\mathbf{x}^\mathsf{T}\mathbf{Y}$. The concentration behavior for these nonlinear quadratic forms is precisely characterized in the following result.

Theorem 1.24 (**Concentration of nonlinear quadratic forms**, [20, 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 $\phi\colon \mathbb{R} \to \mathbb{R}$ with Lipschitz constant K_ϕ and any $t > 0$ that*

$$\mathbb{P}\left(\left|\frac{1}{n}\phi(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\phi(\mathbf{Y}^\mathsf{T}\mathbf{x}) - \frac{1}{n}\operatorname{tr}\mathbf{A}\mathbf{K}_\phi(\mathbf{Y})\right| \ge \frac{t}{\sqrt{n}}\right) \le \exp\left(-\frac{C}{K_\phi^2}\min\left(\frac{t^2}{(|\phi(0)| + K_\phi\sqrt{p/n})^2}, \sqrt{n}t\right)\right), \tag{1.35}$$

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

Theorem 1.24 can be seen as a *nonlinear* extension of the Hanson–Wright inequality in Theorem 1.22. In particular, in the case of $\mathbf{Y} = \mathbf{I}_n$ with $p = n$ and $f(\mathbf{x})$ having zero mean and unit variance entries, Theorem 1.24 reads

$$\mathbb{P}\left(\left|\frac{1}{n}\phi(\mathbf{x})^\mathsf{T}\mathbf{A}\phi(\mathbf{x}) - \frac{1}{n}\operatorname{tr}\mathbf{A}\right| \ge \frac{t}{\sqrt{n}}\right) \le \exp\left(-\frac{C}{K_\phi^2}\min\left(\frac{t^2}{(|\phi(0)| + K_\phi)^2}, \sqrt{n}t\right)\right). \tag{1.36}$$

This is in accordance with the Hanson–Wright inequality in Theorem 1.22, since for Lipschitz $\phi\colon \mathbb{R} \to \mathbb{R}$ of Lipschitz constant K_ϕ and standard Gaussian $\mathbf{x}$, the entries of $\phi(\mathbf{x})$ are sub-gaussian with sub-gaussian norm K_ϕ. Note, however, that in the general case with $\mathbf{Y} \neq \mathbf{I}_n$, $\phi(\mathbf{Y}^\mathsf{T}\mathbf{x})$ does *not* have independent entries, and so Theorem 1.22 does *not* apply, at least directly, to prove Theorem 1.24 for generic $\mathbf{Y}$.

Remark 1.25 (**Concentration of nonlinear quadratic form observation of large random vectors**)**.** Similar to Remark 1.23, it follows from Theorem 1.24 that the nonlinear quadratic observation $\frac{1}{n}\phi(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\phi(\mathbf{Y}^\mathsf{T}\mathbf{x})$, for Lipschitz f, behaves as

$$\frac{1}{n}\phi(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\phi(\mathbf{Y}^\mathsf{T}\mathbf{x}) \simeq \frac{1}{n}\operatorname{tr}\mathbf{A}\mathbf{K}_\phi(\mathbf{Y}) + O(n^{-1/2}), \tag{1.37}$$

for n large, with $\max\{\phi(0), K_\phi, p/n\} = O(1)$. This, again, leads to the first- and second-order behaviors as:

1. in the first order, $\frac{1}{n}\phi(\mathbf{x}^\mathsf{T}\mathbf{Y})\mathbf{A}\phi(\mathbf{Y}^\mathsf{T}\mathbf{x})$ fluctuates around the deterministic quantity $\frac{1}{n}\operatorname{tr}\mathbf{A}\mathbf{K}_\phi(\mathbf{Y})$; and
2. in the second order, it *concentrates* around this deterministic quantity with a fluctuation of order $O(n^{-1/2})$ with a *sub-gaussian* tail when close to the deterministic quantity, and with a *sub-exponential* tail when far away from the deterministic quantity.

1.5 Looking beyond random scalars and vectors

We have seen in Chapters 1.3 and 1.4 that, while large-dimensional random vectors $\mathbf{x} \in \mathbb{R}^n$ themselves do *not* concentrate (see Observation 1.10 and an illustration in Figure 1.1), their (linear, Lipschitz, quadratic, and even nonlinear quadratic) scalar observations establish concentration behavior of the type

$$f(\mathbf{x}) \simeq \mathbb{E}[f(\mathbf{x})] + o(1), \tag{1.38}$$

for some observation map $f\colon \mathbb{R}^n \to \mathbb{R}$, with high probability, for n large, and some small order term $o(1)$ that vanishes as the dimension n grows large. In the aforementioned examples of linear, Lipschitz, quadratic, and nonlinear quadratic forms, this small $o(1)$ term is shown to be $O(n^{-1/2})$.

Matrices, as natural extension of vectors, are expected to establish similar behaviors. For a large-dimensional random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, one may expect the following.

1. Similar to Observation 1.10 for vectors, the random matrices themselves do *not* concentrate, e.g., in a spectral norm sense, such that $\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\| \not\to 0$ as $n \to \infty$, as we shall below in Theorem 5.7.

2. At the same time, extending the scalar observation maps of vectors, in Definition 1.17, a similar large-dimensional *concentration* behavior for the scalar (e.g., eigenspectral) observations $f(\mathbf{X})$ of the random matrix $\mathbf{X}$ can be observed for certain *matrix functionals* $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}$ of $\mathbf{X}$.

As such, when one is interested *only* in scalar observations—in the matrix case, this could correspond to "trace queries," "quadratic form queries," or other (eigenspectral) functionals $f(\cdot)$ of a random matrix $\mathbf{X}$ that return a scalar (these are common operations of interest in ML and beyond)—then it is often possible to find a deterministic matrix $\bar{\mathbf{X}}$ that "mimics" the behavior of $\mathbf{X}$ but *only* through the observation map $f(\cdot)$.

We refer to such a matrix $\bar{\mathbf{X}}$ as a *Deterministic Equivalent* of $\mathbf{X}$.

(This is a special case of the High-Dimensional Equivalent in Definition 1.1, and it will be formally defined in Definition 6.1 of Chapter 6.) For this Deterministic Equivalent, for any appropriate scalar observation function $f : \mathbb{R}^{p\times n} \to \mathbb{R}$ of $\mathbf{X}$, we have, akin to Equation (1.38), that

$$f(\mathbf{X}) \simeq f(\bar{\mathbf{X}}) + o(1), \tag{1.39}$$

for n, p large. As a consequence, these scalar observations $f(\bar{\mathbf{X}})$ of Deterministic Equivalents "track" the behavior of their random equivalent $f(\mathbf{X})$ with increased accuracy as the dimensions n, p grow large. In Chapter 6 of Part II, we will showcase both types of results using the example of sample covariance matrix. For that purpose, we will need a few basic linear algebraic notations and results. These will be reviewed in the next chapter.

Chapter 2

Basic linear algebra

In this chapter, we briefly review basic linear algebraic notations and results to be used throughout the monograph. In Chapter 2.1, we review inner products and norm of vectors and matrices in the Euclidean space. These results, when combined with probabilistic arguments discussed in Chapter 1, provide novel insights into classical linear algebraic statements, for both vectors and matrices. As an example, we shall see in Chapter 2.2 that matrix norms are *not* so equivalent for matrices of large size. In Chapter 2.3, we recall spectral (i.e., eigenvalue and singular value) decompositions of matrices. Finally, in Chapter 2.4, we describe connection between spectral decompositions and solving linear equations.

2.1 Inner products and norms for vectors and matrices

Vectors. The inner product and the related notions of Euclidean norm, angle, and orthogonality are among the most basic quantities that are widely used to describe properties of vectors.

Definition 2.1 (**Inner product, Euclidean norm, angle, and orthogonality**). *Given vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *living in the* n*-dimensional Euclidean space* $\mathbb{R}^n$ *composed as* $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$ *and* $\mathbf{y} = [y_1, \ldots, y_n]^\mathsf{T}$*, respectively,*

1. $\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_{i=1}^n x_i y_i$ *is the* inner product *between* $\mathbf{x}$ *and* $\mathbf{y}$*;*
2. $\|\mathbf{x}\|_2^2 = \mathbf{x}^\mathsf{T}\mathbf{x} = \sum_{i=1}^n x_i^2$ *is the* (squared) Euclidean norm *of* $\mathbf{x}$*; and*
3. $\cos\theta = \left(\frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2}\right)$ *is the* (cosine of the) angle *between* $\mathbf{x}$ *and* $\mathbf{y}$.

We say that the vectors $\mathbf{x}, \mathbf{y}$ *are* orthogonal *to each other if* $\mathbf{x}^\mathsf{T}\mathbf{y} = 0$*; in this case,* $\cos\theta = 0$*, and* $\theta = \pi/2$.

Remark 2.2 (**Vector Euclidean norm as a total energy/mass**). Intuitively, the Euclidean norm $\|\mathbf{x}\|_2$ measures the total "mass" or "energy" of the vector $\mathbf{x} \in \mathbb{R}^n$, and this can be decomposed in various ways. Somewhat more precisely, for $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n$, the canonical vectors of $\mathbb{R}^n$ with $[\mathbf{e}_i]_j = \delta_{ij}$ (that, in particular, form an orthonormal basis of $\mathbb{R}^n$), any $\mathbf{x} \in \mathbb{R}^n$ admits the following decomposition

$$\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i, \tag{2.1}$$

with x_i the i^{th} entry of $\mathbf{x}$. It follows that $\|\mathbf{x}\|_2^2$ collects (the squared sum of) all the entries x_i, i.e., $\|\mathbf{x}\|_2^2 = \sum_{i=1}^n x_i^2$. This is a generalization of the Pythagorean theorem.

The Euclidean norm of vectors in $\mathbb{R}^n$ can be defined (as it was in Definition 2.1) in terms of inner products. The converse statement, that inner products can be characterized in terms of norms, is also true. It is known as a *polarization identity*, and it is given in the following result.

Lemma 2.3 (**Polarization identity**). *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have*

$$\mathbf{x}^\mathsf{T}\mathbf{y} = \frac{1}{2}\left(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right). \tag{2.2}$$

Lemma 2.3 connects the inner product $\mathbf{x}^\mathsf{T}\mathbf{y}$ to the Euclidean norm of the distance between $\mathbf{x}$ and $\mathbf{y}$, $\|\mathbf{x}-\mathbf{y}\|_2^2$, as well as the Euclidean norms of $\mathbf{x}$ and $\mathbf{y}$. The value of $\mathbf{x}^\mathsf{T}\mathbf{y}$ can be positive or negative, depending on whether the two vectors are in the same "direction" or not.

Polarization identities are usually presented simply as linear algebraic facts for given vectors $\mathbf{x}$ and $\mathbf{y}$. However, when combined with a probabilistic modeling for $\mathbf{x}$ (and/or $\mathbf{y}$) as a random vector living in $\mathbb{R}^n$, Lemma 2.3 can be used to provide an explanation for a counterintuitive behavior of large-dimensional random (data) vectors. This is illustrated in **??** and **??**; and it is discussed in the following remark.

Remark 2.4 (**Polarization identity and different scaling for inner products and norms of large random vectors**). For fixed vector $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\|_2 = 1$ and *random* vector $\mathbf{x} \in \mathbb{R}^n$ such that $\sqrt{n}\mathbf{x}$ has i.i.d. entries with zero mean, unit variance, and finite fourth order moment $m_4 < \infty$ (the scaling by $\sqrt{n}$ is made so that $\mathbb{E}[\|\mathbf{x}\|_2^2] = 1$), we have the following.

1. It follows from the LLN and CLT (in Theorems 1.7 and 1.8, respectively) that

$$\mathbf{x}^\mathsf{T}\mathbf{y} \simeq 0 + \mathcal{N}(0,1)/\sqrt{n}, \tag{2.3}$$

   for n large, so that the (random) inner product $\mathbf{x}^\mathsf{T}\mathbf{y}$ is of order $O(n^{-1/2})$ with high probability.

2. On the other hand, again by the LLN, CLT, and the fact $\mathbb{E}[(\mathbf{x}^\mathsf{T}\mathbf{x})^2] = \frac{n+m_4-1}{n}$, one has that

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^\mathsf{T}\mathbf{x} \simeq 1 + \mathcal{N}(0, m_4 - 1)/\sqrt{n}, \tag{2.4}$$

   for large n, so that the (random) Euclidean norm $\|\mathbf{x}\|_2 \simeq 1$, and thus is of order $O(1)$.

3. It then follows from the Polarization identity in Lemma 2.3 that

$$\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + O(n^{-1/2}) = 2 + O(n^{-1/2}), \tag{2.5}$$

   for large n, so that the Euclidean distance between $\mathbf{x}$ and *any* fixed $\mathbf{y}$ (or their norms) is much larger (in fact by a factor of $\sqrt{n}$) than their inner product.

Recall from Definition 2.1 that two vectors $\mathbf{x}, \mathbf{y}$ are orthogonal if $\mathbf{x}^\mathsf{T}\mathbf{y} = 0$. Thus, by Remark 2.4, one has that a large-dimensional random vector $\mathbf{x}$ having i.i.d. entries is always *approximately orthogonal* to any deterministic vector $\mathbf{y}$. This is also a manifestation of the "non-concentration" (or CLT-type concentration) behavior of large-dimensional random vectors discussed in Observation 1.10 and illustrated in Figure 1.1. This intrinsically different scaling (by $\sqrt{n}$) between the norm and inner-product/angle of large-dimensional random vectors comes from the fundamental concentration behavior (e.g., LLN and CLT in Theorems 1.7 and 1.8); and it will, as we shall see below in Chapter 3, distinguish the two regimes of interest for nonlinear (random) functions.

One can consider other vector norms beyond the Euclidean norm.

Definition 2.5 (**p-norm of vectors**). *For any real number $p \geq 1$ and $\mathbf{x} \in \mathbb{R}^n$, the p-norm (also known as the ℓ_p norm) of $\mathbf{x}$ is defined as*

$$\|\mathbf{x}\|_p \equiv \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{2.6}$$

As a special case, we obtain the Manhattan norm *with $p = 1$, the* Euclidean norm *with $p = 2$, and the* infinity/maximum *norm with $p \to \infty$ as $\|\mathbf{x}\|_\infty \equiv \max_i |x_i|$.*

Remark 2.6 (**Vector norm "equivalence"**). For a vector $\mathbf{x} \in \mathbb{R}^n$, one has

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2 \leq n\|\mathbf{x}\|_\infty, \tag{2.7}$$

so that the vector norms in Definition 2.5 are "equivalent," but only up to a factor that depends on the the dimension n.

Matrices. The previous results hold for vectors, but they generalize very naturally to matrices. For matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$, we can use the matrix trace function to define their inner product and associated *Frobenius norm*, as follows.

Definition 2.7 (**Matrix inner product and Frobenius norm**). *Given matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$,*

1. $\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{Y}) = \sum_{i=1}^{n} [\mathbf{X}^\mathsf{T}\mathbf{Y}]_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ji} Y_{ji}$ *is the* matrix inner product *between $\mathbf{X}$ and $\mathbf{Y}$, where* $\mathrm{tr}(\mathbf{A})$ *is the trace of $\mathbf{A}$ (that is also equal to the sum of all eigenvalues* and *diagonal entries of $\mathbf{A}$, see Definition 2.19 below); and*

2. $\|\mathbf{X}\|_F^2 = \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) = \sum_{i=1}^{n} [\mathbf{X}^\mathsf{T}\mathbf{X}]_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ji}^2$ *denotes the* (squared) Frobenius norm *of $\mathbf{X}$ (that is also equal to the sum of the squared entries of $\mathbf{X}$).*

As with vectors, we have polarization identities and (when combined with a probabilistic modeling for the elements of the matrices) associated scaling considerations for matrices.

Remark 2.8 (**Polarization identity and different scaling for large random matrices**). Similar to Lemma 2.3, for matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m \times n}$, we have the following matrix polarization identity:

$$\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{Y}) = \frac{1}{2} \left(\|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 - \|\mathbf{X} - \mathbf{Y}\|_F^2 \right). \tag{2.8}$$

Also, similar to Remark 2.4, we have for fixed $\mathbf{Y}$ with $\|\mathbf{Y}\|_F = 1$ and random $\mathbf{X} \in \mathbb{R}^{m \times n}$ such that $\sqrt{mn}\ \mathbf{X}$ has i.i.d. entries of zero mean, unit variance, and finite fourth order moment $m_4 < \infty$ (again, the scaling $\sqrt{mn}$ is made so that $\mathbb{E}[\|\mathbf{X}\|_F^2] = 1$) that

$$\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{Y}) \simeq 0 + \mathcal{N}(0, 1)/\sqrt{mn}, \tag{2.9}$$

and

$$\|\mathbf{X}\|_F^2 \simeq 1 + \mathcal{N}(0, m_4 - 1)/\sqrt{mn}, \tag{2.10}$$

so that

$$\|\mathbf{X} - \mathbf{Y}\|_F^2 = \|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 + O(1/\sqrt{mn}) = 2 + O(1/\sqrt{mn}). \tag{2.11}$$

That is, Remark 2.4 extends naturally to matrices.

As with vector norms, there are many widely-used matrix norms beyond the Frobenius norm. One class of such matrix norms is discussed as follows.

Definition 2.9 (**Matrix norm**). *For $\mathbf{X} \in \mathbb{R}^{p \times n}$, consider the following "entry-wise" extension of the p-norms of vectors in Definition 2.5:*

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.*

Also, we can consider the matrix norm induced by vectors, defined as

$$\|\mathbf{X}\|_p \equiv \sup_{\|\mathbf{v}\|_p=1} \|\mathbf{X}\mathbf{v}\|_p. \tag{2.12}$$

By taking $p = 2$ *in Equation* (2.12) *we get the* spectral norm*, defined as*

$$\|\mathbf{X}\|_2 = \sqrt{\lambda_{\max}(\mathbf{X}\mathbf{X}^\mathsf{T})} = \sigma_{\max}(\mathbf{X}),$$

where $\lambda_{\max}(\mathbf{X}\mathbf{X}^\mathsf{T})$ *and* $\sigma_{\max}(\mathbf{X})$ *denotes the maximum eigenvalue and singular of* $\mathbf{X}\mathbf{X}^\mathsf{T}$ *and* $\mathbf{X}$*, respectively.*

The matrix Frobenius norm and spectral norm in Definition 2.9 belong to the class of so-called matrix *Schatten norms* (that can be defined by applying the vector p-norms in Definition 2.5 on the vector of *singular values* of the matrix). These norms are known to be *unitarily invariant*, i.e., such that $\|\mathbf{X}\| = \|\mathbf{U}\mathbf{X}\mathbf{V}\|$ for all matrices $\mathbf{X}$ and unitary (square) matrices $\mathbf{U}, \mathbf{V}$ of appropriate dimensions.

We have the following inequalities between different matrix norms that establish a certain sort of equivalence between matrix norms (that is often too loose for practical use, though).

Remark 2.10 (**Matrix norm "equivalence"**)**.** For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, one has the following

1. $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{\mathrm{rank}(\mathbf{A})} \cdot \|\mathbf{A}\|_2 \leq \sqrt{\max(m,n)} \cdot \|\mathbf{A}\|_2$; so that, e.g., the control of the Frobenius norm via the spectral norm can be particularly loose for matrices of large rank and/or size; 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 norm and spectral norm can be significantly different for large matrices.

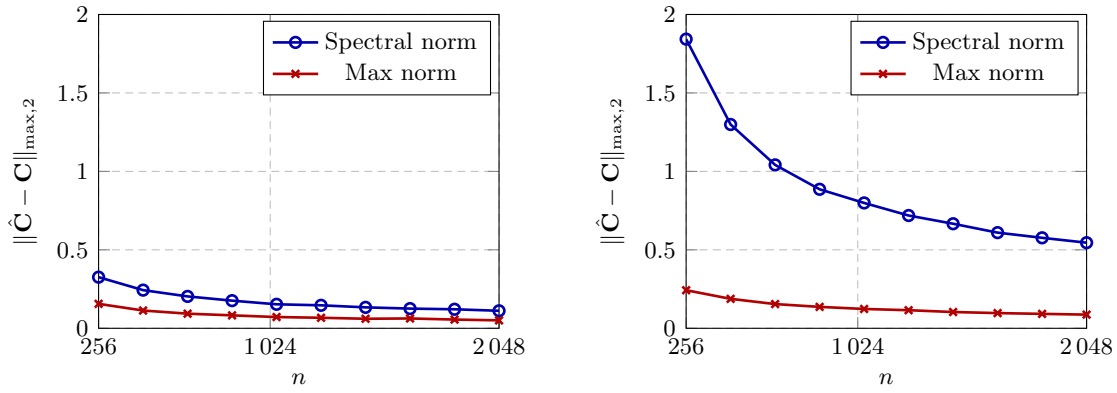
The fact that this notion of matrix norm "equivalence" holds only up to dimensional factors is crucial in large-dimensional data analysis and machine learning. In Chapter 2.2, we will discuss this in more detail with the example of sample covariance matrix, and two popular dimension reduction techniques: Principle Component Analysis (PCA) and multidimensional scaling (MDS).

2.2 Loss of matrix norm equivalence in ML

In this section, we delve further into the "(loss of) matrix norm equivalence" discussed in Remark 2.10, using the sample covariance matrix (SCM) in the proportional regime as an illustrative example.[2][3] Then, we discuss how this "(loss of) matrix norm equivalence" has a significant impact on large-scale ML, with the examples of two popular dimension reduction techniques: principle component analysis (PCA, that is directly connected to SCM) and multidimensional scaling (MDS), in Example 2.14 and Example 2.17, respectively.

[2]For the formal definitions of SCM and proportional regime, see Definitions 4.22 and 4.23, respectively.

[3]We assume basic familiarity with eigenvalues/eigenvectors; these are described in more detail in Chapter 2.3.



(a) $p = 8$ (b) $p = 128$

Figure 2.1: Spectral versus max norm errors of $\hat{\mathbf{C}} - \mathbf{C}$, as a function of the sample size n, for $p = 8$ (Figure 2.1a) and $p = 128$ (Figure 2.1b), with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and thus $\mathbf{C} = \mathbf{I}_p$. Results averaged over 50 independent runs.

Example 2.11 (**Loss of matrix norm equivalence for SCM**)**.** Consider a set of n independent random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ following a multi-variate Gaussian distribution, with zero mean and identity covariance, $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}_p$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. In this case, the SCM is given by

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T}. \tag{2.13}$$

This quantity is known to be the maximum likelihood estimator of the population covariance $\mathbf{C} = \mathbf{I}_p$, and thus it should be the "optimal" solution we can get.

Now, we evaluate the maximum and spectral norm (see Definition 2.9 above) of the SCM $\hat{\mathbf{C}}$ in the proportional regime, by considering the limit of $n, p \to \infty$ with $p/n \to c \in (0, \infty)$. In this setting, we have the following dual observations.

1. From the LLN in Theorem 1.7, it follows that that the (i, j) entry of the SCM $\hat{\mathbf{C}}$ converges to the population covariance $\mathbf{C} = \mathbf{I}_p$ as $n \to \infty$. That is,
$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_{\max} \to 0. \tag{2.14}$$

2. On the other hand, if we let $n, p \to \infty$ with $p > n$, then $\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ in Equation (2.13) is the sum of n rank-one matrices, and the rank of $\hat{\mathbf{C}}$ is *at most* equal to n. In this case, being a $p \times p$ matrix with $p > n$, the sample covariance matrix $\hat{\mathbf{C}}$ must be a *singular* matrix having at least $p - n > 0$ zero eigenvalues. As a consequence of this *eigenvalue mismatch*, we have
$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \neq 0, \tag{2.15}$$
as long as $p > n$, even for n, p *arbitrary large* as $n, p \to \infty$.

While the eigenvalue mismatch in Equation (2.15) may, at first sight, seem to contradict the max norm convergence results in Equation (2.14), this is not the case. This is a consequence of the fact that matrix norms are "equivalent," but only up to factors that depend on the size p of the matrix, as already mentioned in Remark 2.10.[4] For instance, we have

$$\|\mathbf{A}\|_{\max} \leq \|\mathbf{A}\|_2 \leq p \|\mathbf{A}\|_{\max} \tag{2.16}$$

for the symmetric matrix $\mathbf{A} = \hat{\mathbf{C}} - \mathbf{I}_p \in \mathbb{R}^{p \times p}$. The conclusion is that, when considering statistical problems of large dimensions (with $p \gg 1$), the proportional regime:

[4] That is, the sense in which "all matrix norms are equivalent" is very weak, depending on dimensional factors that have strong algorithmic and statistical consequences, the latter being of particular interest for RMT.

matrix norms are *not* equivalent in the proportional regime.

As an illustration of this loss of matrix matrix norm equivalence in the proportional regime, Figure 2.1 provides numerical evidences on the errors in max norm and spectral norm of $\hat{\mathbf{C}} - \mathbf{I}_p$ in the classical ($p = 8$ in Figure 2.1a) and proportional ($p = 128$ in Figure 2.1b) regimes. In particular, we see that the differences between the spectral and max norms are *significantly smaller* in the classical regime than in the proportional regime. In the proportional regime, with $p = 128$ (which is indeed *not* very large in the context of modern ML), the relative error in spectral norm can blow up to 200%, while the max norm error still remains at a much lower level (of less than 25%).

Thus, control on the max norm does *not* yield, at least directly, a non-trivial control on the spectral norm that is often of more practical interest in ML. The practical usefulness of the spectral norm is discussed in more detail in the following remark.

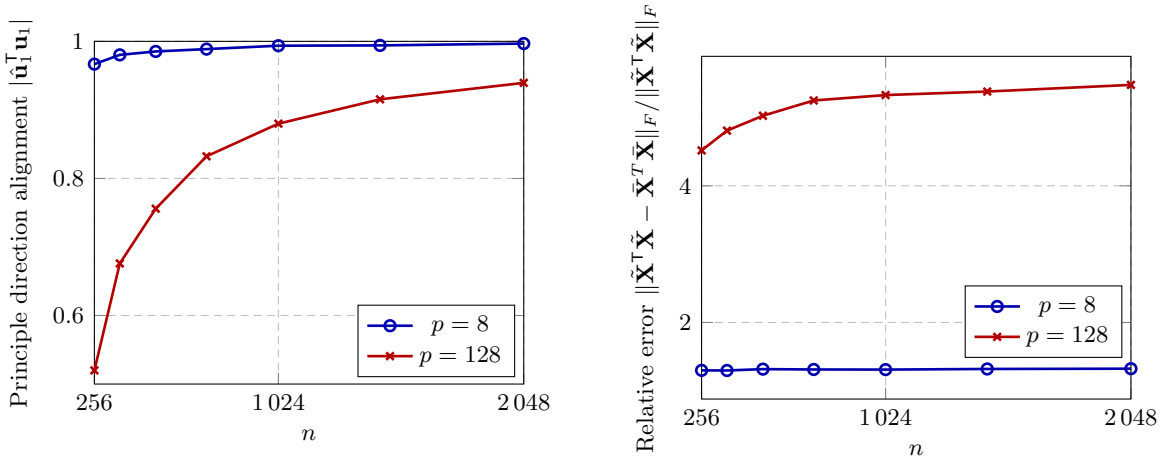
Remark 2.12 (**On the importance of spectral norm**)**.** For practical purposes, this "loss of norm equivalence" for large matrices (large p) raises the question of what is the relevant matrix norm to consider for a given problem. For many ML problems, the spectral norm is the most relevant, in the following sense.

1. First, the spectral norm is the matrix norm induced by the Euclidean norm of vectors (see for example [18, Theorem 5.6.2]). Thus, the study of regression vectors or label/score vectors in classification is naturally attached to the eigenspectral study of matrices. (See the problem of linear least squares regression in **??** as an instance of this.)

2. Second, one needs to evaluate the spectral norm when spectral methods such as principle component analysis (PCA) [40], multi-dimensional scaling (MDS) [41], (kernel) spectral clustering [25] or PageRank [13] are considered. More precisely, for matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$, according to Weyl's inequality (see [18, Theorem 4.3.1] and Lemma A.3 in Appendix A), one has
$$\max_{1\le i\le p} |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})\|_2 \le \|\mathbf{A} - \mathbf{B}\|_2, \tag{2.17}$$
for $\lambda_1(\mathbf{A}) \ge \lambda_2(\mathbf{A}) \ge \ldots \ge \lambda_p(\mathbf{A})$ the eigenvalues of $\mathbf{A}$ in a decreasing order. Thus, the bound on the spectral norm difference provides a uniform bound on *all* the corresponding eigenvalues. See Also, it follows from Davis–Kahan theorem (see [7] and Lemma A.4 in Appendix A) that
$$\sqrt{1 - (\mathbf{u}_i^\mathsf{T}(\mathbf{A})\mathbf{u}_i(\mathbf{B}))^2} \le \frac{\|\mathbf{A} - \mathbf{B}\|_2}{\min\{|\lambda_{i-1}(\mathbf{A}) - \lambda_i(\mathbf{B})|, |\lambda_{i+1}(\mathbf{A}) - \lambda_i(\mathbf{B})|\}} \tag{2.18}$$
for $\mathbf{u}_i(\mathbf{A}), \mathbf{u}_i(\mathbf{B})$ the eigenvector that corresponds to the eigenvalue of $\lambda_i(\mathbf{A})$ and $\lambda_i(\mathbf{B})$, respectively. Thus, the "alignment" between corresponding eigenvectors and subspaces can be controlled by the spectral norm. See Example 2.14 below for an application Principle Component Analysis and **??** for an application to spectral clustering.[5]

Definition 2.13 (**Principle component analysis**, **PCA**)**.** *For data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ of dimension p, denote its SCM $\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\mathsf{T} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ as in Example 2.11, PCA aims to find the* principle direction $\mathbf{u} \in \mathbb{R}^p$ *of $\mathbf{X}$ by solving the following optimization problem*
$$\begin{aligned} \max_{\mathbf{u}\in\mathbb{R}^p} \quad & \mathbf{u}^\mathsf{T}\hat{\mathbf{C}}\mathbf{u} \\ s.t. \quad & \|\mathbf{u}\| = 1. \end{aligned} \tag{2.19}$$

[5] Most previous literature on RMT has been concerned with the eigenvalues of random matrices and functional of them. In ML applications, however, eigenvectors are more commonly exploited and thus of more practical interest. Technically speaking, to characterize the eigenvectors one needs to the evaluate the behavior of the *whole* random matrix (instead of solely its eigenvalues). This can be achieved with the proposed Deterministic Equivalent for resolvent framework, to be discussed in Chapter 6.



(a) PCA (b) MDS

Figure 2.2: **Figure 2.2a**: Principle direction alignment $|\hat{\mathbf{u}}_1^\mathsf{T}\mathbf{u}_1|$ for $\hat{\mathbf{u}}_1$ the principle direction obtained from PCA; and **Figure 2.2b**: related approximation error in Frobenius norm $\|\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} - \bar{\mathbf{X}}^T\bar{\mathbf{X}}\|_F / \|\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}}\|_F$ obtained from MDS with $m = 1$; as a function of the sample size n, for $p = 8$ (**red**) and $p = 128$ (**blue**), with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ for $\mathbf{C} = \mathbf{I}_p + \mathbf{u}_1\mathbf{u}_1^\mathsf{T}$ and $\mathbf{u}_1 = [\mathbf{1}_{p/2}; -\mathbf{1}_{p/2}]/\sqrt{p}$. Results averaged over 50 independent runs.

Denote $\hat{\mathbf{C}} = \mathbf{U}_{\hat{\mathbf{C}}}\mathbf{\Lambda}_{\hat{\mathbf{C}}}\mathbf{U}_{\hat{\mathbf{C}}}^\mathsf{T}$ *the eigen-decomposition (see Definition 2.19 below for a formal definition) of* $\hat{\mathbf{C}}$*, for diagonal* $\mathbf{\Lambda}_{\hat{\mathbf{C}}} = \mathrm{diag}\{\lambda_i(\hat{\mathbf{C}})\}_{i=1}^p$ *containing the eigenvalues of* $\hat{\mathbf{C}}$ *and orthonormal* $\mathbf{U}_{\hat{\mathbf{C}}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_n] \in \mathbb{R}^{p\times p}$ *containing the corresponding eigenvectors. Then, the top eigenvector* $\hat{\mathbf{u}}_1 \in \mathbb{R}^p$ *that corresponds to the largest eigenvalue* $\lambda_1(\hat{\mathbf{C}})$ *is the solution to* (2.19)*. This is the "direction" where the data distribution is the most extended, and it "explains" most of the variability in the data.*

Subsequent principle directions/components of the data can be similarly computed. Denote $\mathbf{U}_{\hat{\mathbf{C}},m} \in \mathbb{R}^{p\times m}$ *the* m*-principle components of* $\mathbf{X} \in \mathbb{R}^{p\times n}$*, an* m*-dimensional representations of* $\mathbf{X}$ *obtained from PCA is given by (the columns of)* $\tilde{\mathbf{X}} = \mathbf{U}_{\hat{\mathbf{C}},m}^\mathsf{T}\mathbf{X} \in \mathbb{R}^{m\times n}$*.*

Example 2.14 (**Principle component analysis in high dimensions**)**.** As a consequence of the loss of SCM norm equivalence in Example 2.11 and the importance of spectral norm in Remark 2.12, we should *not*, a prior, expect that the popularly used PCA dimension reduction approach described in Definition 2.13 works well for large-dimensional data vectors. Figure 2.2a provides numerical illustrations of the different behavior of PCA in the classical versus proportional regime. For i.i.d. multi-variate Gaussian data vector $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ with covariance $\mathbf{C} = \mathbf{I}_p + \mathbf{u}_1\mathbf{u}_1^\mathsf{T}$, we evaluate here the "alignment" $|\hat{\mathbf{u}}_1^\mathsf{T}\mathbf{u}_1|$ between the principle direction $\hat{\mathbf{u}}_1$ obtained from SCM and the true covariance principle direction $\mathbf{u}_1 \in \mathbb{R}^p$, for small $p = 8$ and large $p = 128$, and sample size n ranging from 256 to 2 048. From Figure 2.2a, we see that while for p small, the principle direction $\hat{\mathbf{u}}_1$ obtained from PCA constantly aligns to the true data principle direction, this is no longer the case for p large. This is a consequence of the (now uncontrolled, for p large) spectral norm different $\|\hat{\mathbf{C}} - \mathbf{C}\|_2$.

Another commonly used dimension reduction technique is multidimensional scaling (MDS) [41]. Different from PCA in Example 2.14, classical MDS aims to obtain low-dimensional (in $\mathbb{R}^m$ say) representation of the data so that their Euclidean distances (or dissimilarities) are approximately preserved. This is described as follows.

Definition 2.15 (**Multidimensional scaling**, **MDS**)**.** *Classical MDS aims to obtain low-dimensional (in* $\mathbb{R}^m$ *say) representation of the data so that their Euclidean distances (or dissimilarities) are approximately preserved. More precisely, for data vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ *of*

dimension p, denote $\mathbf{E} \equiv \{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/p\}_{i,j=1}^n$ their (normalized) squared Euclidean distance matrix, MDS aims to find m-dimensional representations $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n] \in \mathbb{R}^{m\times n}$ of $\mathbf{X}$ such that their Euclidean distances are approximately preserved:

$$\tilde{\mathbf{E}} \equiv \{\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2^2/p\}_{i,j=1}^n \approx \mathbf{E}. \tag{2.20}$$

To do this, note that

$$\mathbf{E} = \mathbf{1}_n\mathbf{v}^\mathsf{T} + \mathbf{v}\mathbf{1}_n^\mathsf{T} - 2\mathbf{X}^\mathsf{T}\mathbf{X}, \quad \mathbf{v} = \{\|\mathbf{x}_i\|^2\}_{i=1}^n, \tag{2.21}$$

so that by performing "double centering" of $\mathbf{E}$ we get $-\frac{1}{2}\mathbf{PEP} = \mathbf{PX}^\mathsf{T}\mathbf{XP} = \bar{\mathbf{X}}^\mathsf{T}\bar{\mathbf{X}}$, for $\bar{\mathbf{X}} = \mathbf{XP}$ with $\mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}$. Then, $\tilde{\mathbf{X}}$ is obtained by minimizing the following strain*:*

$$\min_{\tilde{\mathbf{X}}\in\mathbb{R}^{m\times n}} \quad \|\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} - \bar{\mathbf{X}}^\mathsf{T}\bar{\mathbf{X}}\|_F^2. \tag{2.22}$$

Per the Eckart-Young-Mirsky theorem [12, 24], the solution to (2.22) *is given by $\tilde{\mathbf{X}} = \mathbf{\Sigma}_{\bar{\mathbf{X}},m}\mathbf{V}_{\bar{\mathbf{X}},m}^\mathsf{T}$, for $\bar{\mathbf{X}} = \mathbf{U}_{\bar{\mathbf{X}}}\mathbf{\Sigma}_{\bar{\mathbf{X}}}\mathbf{V}_{\bar{\mathbf{X}}}^\mathsf{T}$ the singular value decomposition of $\bar{\mathbf{X}} \in \mathbb{R}^{p\times n}$, and $\mathbf{V}_{\bar{\mathbf{X}},m} \in \mathbb{R}^{n\times m}$ and $\mathbf{\Sigma}_{\bar{\mathbf{X}},m} \in \mathbb{R}^{m\times m}$ containing the top-m right singular vectors and singular values, respectively.*

Remark 2.16 (PCA and MDS)**.** MDS is similar to PCA in Definition 2.13, in that they both provide low-dimensional representation $\tilde{\mathbf{X}} \in \mathbb{R}^{m\times p}$ of the data $\mathbf{X} \in \mathbb{R}^{p\times n}$ with $m \ll p$. From an algorithmic aspect, they are connected to each other through the singular value decomposition (SVD, see Definition 2.22 for a formal definition) of $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$ and of $\mathbf{XP} = \bar{\mathbf{X}} = \mathbf{U}_{\bar{\mathbf{X}}}\mathbf{\Sigma}_{\bar{\mathbf{X}}}\mathbf{V}_{\bar{\mathbf{X}}}^\mathsf{T}$ as follows.

1. By Definition 2.13, PCA computes $\tilde{\mathbf{X}} = \mathbf{U}_{\mathbf{X};m}^\mathsf{T}\mathbf{X} = \begin{bmatrix}\mathbf{I}_m & \mathbf{0}\end{bmatrix}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$, where $\mathbf{U}_{\mathbf{X},m} \in \mathbb{R}^{p\times m}$ is the top-m left singular subspace of $\mathbf{X}$.

2. On the other hand, by Definition 2.15, MDS computes $\tilde{\mathbf{X}} = \mathbf{\Sigma}_{\bar{\mathbf{X}},m}\mathbf{V}_{\bar{\mathbf{X}},m}^\mathsf{T} = \begin{bmatrix}\mathbf{I}_m & \mathbf{0}\end{bmatrix}\mathbf{\Sigma}_{\bar{\mathbf{X}}}\mathbf{V}_{\bar{\mathbf{X}}}^\mathsf{T}$ of the "centered" data matrix $\bar{\mathbf{X}} = \mathbf{XP}$.

As such, classical MDS boils down, up centering and per (2.22), to the evaluation of data Gram matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$, and then to the computation of the data (top) singular values and vectors, for which a similar loss of norm equivalence as for PCA in Example 2.14 is expected. This is discussed as follow.

Example 2.17 (**Multidimensional scaling in high dimensions**)**.** It can be checked that the MDS approximation error in Equation (2.22) is given by the sum of eigenvalues (excluding the largest m) of $\bar{\mathbf{X}}^\mathsf{T}\bar{\mathbf{X}}$ (that coincide with the sum of those of the centered SCM $\bar{\mathbf{X}}\bar{\mathbf{X}}^\mathsf{T}$). Thus, by Remark 2.12, we have, similar to Example 2.14 for PCA and as a consequence of the loss of SCM norm equivalence, that we should *not* expect that the MDS works well for large-dimensional data vectors. See Figure 2.2b for a numerical manifestation of this fact. We particularly see from Figure 2.2b that unlike for $p = 8$, where the relative approximation error is small; in the case of large-dimensional data with $p = 128$, the approximation error is much larger, and increases as n grows large.

2.3 Spectral decomposition of matrices

Here, we review in more detail the spectral decomposition (including both the eigenvalue decomposition and the singular value decomposition) of matrices.

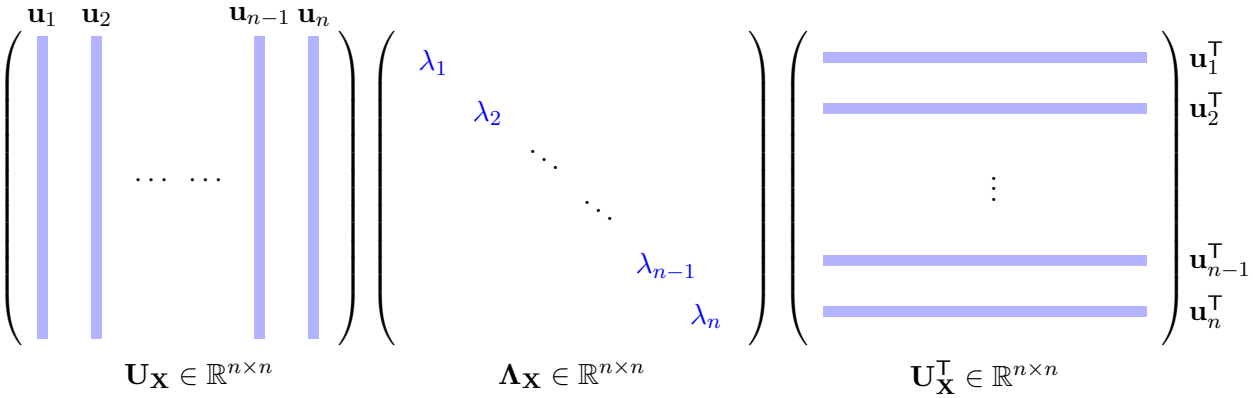


Figure 2.3: Illustration of eigen-decomposition, $\mathbf{U}_\mathbf{X}\mathbf{\Lambda}_\mathbf{X}\mathbf{U}_\mathbf{X}^\mathsf{T}$, of a symmetric $n \times n$ matrix $\mathbf{X}$.

Symmetric and Hermitian matrices. Let's start by recalling the definition and properties of symmetric real and Hermitian complex matrices, as follows.

Definition 2.18 (Symmetric and Hermitian matrix). *For a real square matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, we say $\mathbf{X}$ is symmetric if $\mathbf{X}^\mathsf{T} = \mathbf{X}$. Similarly, for a complex square matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$, we say $\mathbf{X}$ is Hermitian if $\mathbf{X}^* = \mathbf{X}$ (with $\mathbf{X}^*$ the conjugate transpose of $\mathbf{X}$).*

Important facts about symmetric/Hermitian matrices are the following.

1. $\mathbf{X}$ is symmetric *if and only if* there exists real orthonormal $\mathbf{U} \in \mathbb{R}^{n\times n}$ and real diagonal $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ such that $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$.

2. $\mathbf{X}$ is Hermitian *if and only if* there exists unitary $\mathbf{U} \in \mathbb{C}^{n\times n}$ and real diagonal $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ such that $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$.

In more detail, for symmetric real (or Hermitian complex) matrices, their diagonalization leads to the following eigen-decomposition (according to the eigenvalues and eigenvectors of the matrix of interest).

Definition 2.19 (Eigen-decomposition of symmetric matrices, [18, Theorem 2.5.6]). *For a symmetric real matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$ of $\mathbf{X}$ are all real, and $\mathbf{X}$ admits the following eigen-decomposition*

$$\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Lambda}_\mathbf{X}\mathbf{U}_\mathbf{X}^\mathsf{T} = \sum_{i=1}^{n} \lambda_i(\mathbf{X})\mathbf{u}_i\mathbf{u}_i^\mathsf{T}, \tag{2.23}$$

for diagonal $\mathbf{\Lambda}_\mathbf{X} = \mathrm{diag}\{\lambda_i(\mathbf{X})\}_{i=1}^n$ containing the 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, the eigenvalue and eigenvector pair $(\lambda_i(\mathbf{X}), \mathbf{u}_i)$ of $\mathbf{X}$ satisfies the following equation

$$\mathbf{X}\mathbf{u}_i = \lambda_i(\mathbf{X})\mathbf{u}_i. \tag{2.24}$$

See Figure 2.3 for an illustration of eigen-decomposition of symmetric matrix. Given this eigen-decomposition, the *matrix trace* can be defined as $\mathrm{tr}(\mathbf{X}) = \sum_{i=1}^n \lambda_i(\mathbf{X})$.

A similar decomposition as that provided by Definition 2.19 holds for Hermitian complex matrices, by replacing the transpose operators above with conjugate transpose.

In some cases, one is interested in the properties of a single eigenvalue of a symmetric real matrix, $\mathbf{X} \in \mathbb{R}^{n\times n}$. In this case, one may either resort to the eigenvalue-eigenvector equation in (2.24) or to the determinant equation $\det(\mathbf{X} - \lambda\mathbf{I}_n) = 0$.

In other cases, one is interested in the behavior of multiple eigenvalues. In particular, classical RMT is interested in the *joint* behavior of *all* eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X})$. This leads to the definition of the (empirical) *eigenvalue distribution*, or *empirical spectral distribution (ESD)* of $\mathbf{X}$, defined as follows.

Empirical Spectral Distribution (ESD)

Definition 2.20 (**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 $\mathbf{X}$. This can be represented as*

$$\mu_{\mathbf{X}} \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{X})}, \tag{2.25}$$

where δ_x represents the Dirac measure at x.

We note the following important fact regarding the ESD of a symmetric matrix $\mathbf{X}$:

since $\int \mu_{\mathbf{X}}(dx) = 1$, the spectral measure $\mu_{\mathbf{X}}$ of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ (which may be random or not) is a probability measure.

Thus, for $\mu_{\mathbf{X}}$, the ESD of a real symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ of interest, we can talk about the moments of $\mu_{\mathbf{X}}$, just as for scalar random variables, in Definition 1.1. More precisely,

1. $\int t\mu_{\mathbf{X}}(dt) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\mathbf{X})$ is the first moment of $\mu_{\mathbf{X}}$, and it gives the *average* of all eigenvalues of $\mathbf{X}$; and
2. $\int t^2\mu_{\mathbf{X}}(dt) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^2(\mathbf{X})$ 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 $\mathbf{X}$.

An important subset of symmetric and Hermitian matrices is the family of positive-definite and positive semi-definite matrices, defined as below.

Definition 2.21 (**Positive-definite and positive semi-definite matrices, PD and PSD matrices**). *For a real symmetric matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, we say $\mathbf{X}$ is positive-definite if for any nonzero real vector $\mathbf{v} \in \mathbb{R}^n$ we have $\mathbf{v}^\mathsf{T}\mathbf{X}\mathbf{v} > 0$; and we we say $\mathbf{X}$ is positive semi-definite if $\mathbf{v}^\mathsf{T}\mathbf{X}\mathbf{v} \geq 0$. Similarly, for a Hermitian complex matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$, we say $\mathbf{X}$ is positive-definite if for any nonzero complex vector $\mathbf{v} \in \mathbb{C}^n$ we have $\mathbf{v}^*\mathbf{X}\mathbf{v} > 0$; and we we say $\mathbf{X}$ is positive semi-definite if $\mathbf{v}^*\mathbf{X}\mathbf{v} \geq 0$.*

By definition, the eigenvalues of positive-definite matrices are strictly positive, and those of positive semi-definite matrices are non-negative.

General matrices. Going beyond symmetric/Hermitian matrices, non-symmetric real matrices (including, potentially, non-square matrices) generally do not admit an eigen-decomposition, as in Definition 2.19. However, general matrices do admit the following singular value decomposition (SVD).

Definition 2.22 (**Singular value decomposition (SVD), [18, Theorem 2.5.6]**). *For a real and possibly* non-square *matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, the singular values $\sigma_i(\mathbf{X})$ of $\mathbf{X}$ are unique, real and non-negative, and $\mathbf{X}$ admits the following decomposition*

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i(\mathbf{X})\mathbf{u}_i\mathbf{v}_i^\mathsf{T} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}, \tag{2.26}$$

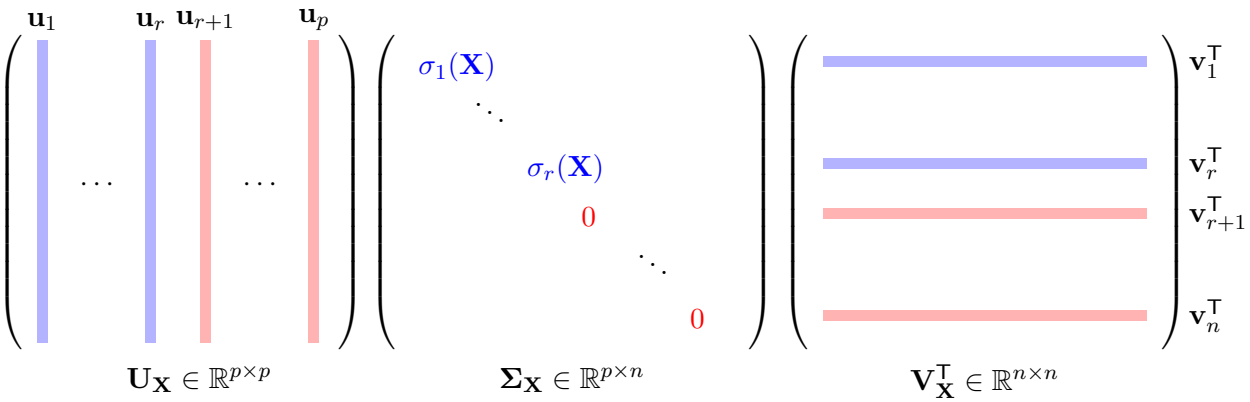


Figure 2.4: Illustration of SVD, $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$, of a $p \times n$ matrix $\mathbf{X}$.

with $r = \mathrm{rank}(\mathbf{X})$*, rectangular diagonal matrix* $\mathbf{\Sigma}_\mathbf{X} \in \mathbb{R}^{p\times n}$ *containing all singular values of* $\mathbf{X}$*, orthonormal* $\mathbf{U}_\mathbf{X} \equiv [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p\times p}$ *and* $\mathbf{V}_\mathbf{X} \equiv [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n\times n}$ *containing the* left *and* right *singular vectors of* $\mathbf{X}$*, respectively. Similar to Equation* (2.24) *for eigenvalue and eigenvector, one has*

$$\mathbf{X}\mathbf{v}_i = \sigma_i(\mathbf{X})\mathbf{u}_i, \quad \mathbf{X}^\mathsf{T}\mathbf{u}_i = \sigma_i(\mathbf{X})\mathbf{v}_i. \tag{2.27}$$

See Figure 2.4 for an illustration of SVD. Similar to Definition 2.20, one may define the empirical distribution of the singular values of a given matrix.

For symmetric positive semi-definite matrices (Definition 2.21), the eigen-decomposition and SVD in Definition 2.19 and Definition 2.22 coincide. Beyond this setting, they are in general different. More generally, however, we have the following connection between the eigen-decomposition and the SVD.

Remark 2.23 (Connection between eigen-decomposition and SVD). For a real matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ with SVD $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$, for orthonormal $\mathbf{U}_\mathbf{X} \in \mathbb{R}^{p\times p}$ and $\mathbf{V}_\mathbf{X} \in \mathbb{R}^{n\times n}$, the eigen-decomposition of $\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{X}^\mathsf{T}\mathbf{X}$ are respectively given by

$$\mathbf{X}\mathbf{X}^\mathsf{T} = \mathbf{U}_\mathbf{X}(\mathbf{\Sigma}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^\mathsf{T})\mathbf{U}_\mathbf{X}^\mathsf{T} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^2\mathbf{U}_\mathbf{X}^\mathsf{T} \in \mathbb{R}^{p\times p}, \tag{2.28}$$

and

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{V}_\mathbf{X}(\mathbf{\Sigma}_\mathbf{X}^\mathsf{T}\mathbf{\Sigma}_\mathbf{X})\mathbf{V}_\mathbf{X}^\mathsf{T} = \mathbf{V}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^2\mathbf{V}_\mathbf{X}^\mathsf{T} \in \mathbb{R}^{n\times n}. \tag{2.29}$$

In this case, the non-zero eigenvalues of $\mathbf{X}\mathbf{X}^\mathsf{T}$ and $\mathbf{X}^\mathsf{T}\mathbf{X}$ are the same (and are the squared singular values of $\mathbf{X}$), and their eigenvectors are connected to the singular vectors of $\mathbf{X}$. More generally, it follows from the Sylvester's determinant theorem (also known as the Weinstein–Aronszajn identity, see Lemma A.9) that for $\mathbf{A} \in \mathbb{R}^{p\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$, one has

$$\det(\mathbf{I}_p + \mathbf{A}\mathbf{B}) = \det(\mathbf{I}_n + \mathbf{B}\mathbf{A}), \tag{2.30}$$

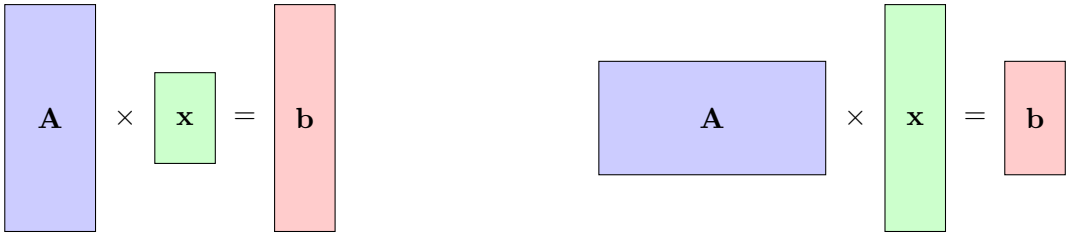
so that the non-zero eigenvalues of $\mathbf{A}\mathbf{B} \in \mathbb{R}^{p\times p}$ and $\mathbf{B}\mathbf{A} \in \mathbb{R}^{n\times n}$ are the same. Also, for a real matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ with SVD $\mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}\mathbf{V}_\mathbf{X}^\mathsf{T}$, we can consider the matrix

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^\mathsf{T} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{(p+n)\times(p+n)}. \tag{2.31}$$

This matrix is real symmetric, and it admits eigen-decomposition

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{U}_\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_\mathbf{X} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_\mathbf{X} \\ \mathbf{\Sigma}_\mathbf{X}^\mathsf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_\mathbf{X}^\mathsf{T} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_\mathbf{X}^\mathsf{T}, \end{bmatrix} \tag{2.32}$$

and that the non-zero singular values of $\mathbf{X}$ are the positive eigenvalues of $\tilde{\mathbf{X}}$.



(a) Over-determined system (b) Under-determined system

Figure 2.5: Illustration of over- versus under-determined linear systems.

2.4 Connection between linear equation and spectral decomposition

One of the most important and fundamental problem in applied mathematics, statistics, and ML is to solve a linear system of equations defined as follow.

Definition 2.24 (**Linear system**). *Given a matrix* $\mathbf{A} \in \mathbb{R}^{p\times n}$ *and a vector* $\mathbf{b} \in \mathbb{R}^p$*, we aim to solve for* $\mathbf{x} \in \mathbb{R}^n$ *that satisfies the following system of linear equations*

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{2.33}$$

For solving linear system, there are in general three regimes of interest.

- *For* $p > n$*, the system has more equations than unknowns; in this case, it is called an* over-determined *system (or sometime, within ML, an* under-parameterized *problem).*
- *When* $p = n$*, the system has the same number of equations and unknowns.*
- *For* $p < n$*, the system has fewer equations than unknowns; in this case, it is called an* under-determined *system (or sometime, within ML, an* over-parameterized *problem).*

See Figure 2.5 for an illustration of the over-determined and under-determined cases.

It should be clear that a solution $\mathbf{x}$ to the linear system in Equation (2.33) exists *if and only if* $\mathbf{b} \in \mathbb{R}^p$ belongs to the column space of $\mathbf{A}$. (That statement is true regardless of the relative sizes of p and n.) In case that there exists a solution, there can be infinitely many, e.g., when the system is under-determined. These solutions can be given using the generalized inverse of $\mathbf{A}$, defined as follows.

Definition 2.25 (**Generalized inverse and Moore–Penrose pseudoinverse**, [15]). *For a real matrix* $\mathbf{A} \in \mathbb{R}^{p\times n}$*, we say the matrix* $\mathbf{A}^g \in \mathbb{R}^{n\times p}$ *is a* generalized inverse *of* $\mathbf{A}$ *if it satisfies*

$$\mathbf{A}\mathbf{A}^g\mathbf{A} = \mathbf{A}. \tag{2.34}$$

Assume, in addition, that the generalized inverse $\mathbf{A}^g$ *satisfies the following additional conditions:*

1. $\mathbf{A}^g\mathbf{A}\mathbf{A}^g = \mathbf{A}^g$*; and*
2. *both* $\mathbf{A}\mathbf{A}^g$ *and* $\mathbf{A}^g\mathbf{A}$ *are symmetric.*

Then, it is the Moore–Penrose pseudoinverse *of* $\mathbf{A}$*, denoted* $\mathbf{A}^+$*.*

A solution to the linear system in Definition 2.24, if it exists, can be fully described using the generalized inverse in Definition 2.25. This is given in the following result.

Theorem 2.26 (**Solution to linear system in Definition 2.24, [15]**). *For* $\mathbf{A}^g \in \mathbb{R}^{n\times p}$*,* any *generalized inverse of* $\mathbf{A}$*, as in Definition 2.25, and the linear system* $\mathbf{A}\mathbf{x} = \mathbf{b}$*, as in Equation* (2.33) *of Definition 2.24,*

1. *the solutions* $\mathbf{x}$ *exist* if and only if $\mathbf{A}\mathbf{A}^g\mathbf{b} = \mathbf{b}$*; and*

2. *all solutions are given by*

$$\mathbf{x} = \mathbf{A}^g\mathbf{b} + (\mathbf{I}_n - \mathbf{A}^g\mathbf{A})\mathbf{w}, \tag{2.35}$$

for arbitrary *vector* $\mathbf{w} \in \mathbb{R}^n$.

In particular, this holds for the Moore–Penrose pseudoinverse $\mathbf{A}^+$ *of* $\mathbf{A}$*. If* $\mathbf{A}$ *has full column rank, then* $\mathbf{I}_n - \mathbf{A}^g\mathbf{A} = \mathbf{0}$*. If* $n = p$ *and* $\mathbf{A}$ *is non-singular, then* $\mathbf{A}^g = \mathbf{A}^{-1}$ *and the solution is unique.*

The generalized inverse $\mathbf{A}^g$ of a matrix $\mathbf{A} \in \mathbb{R}^{p\times n}$, as in Definition 2.25 can be characterized using the SVD of $\mathbf{A}$ in Definition 2.22, as per the following result.

Theorem 2.27 (**Characterization of generalized inverse using SVD**, [15])**.** *Let* $\mathbf{A} \in \mathbb{R}^{p\times n}$ *be a real matrix, with SVD*

$$\mathbf{A} = \mathbf{U}_\mathbf{A} \begin{bmatrix} \mathbf{\Sigma}_\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\mathbf{A}^\mathsf{T}, \tag{2.36}$$

for orthonormal $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{p\times p}$ *and* $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{n\times n}$*, and non-singular* $\mathbf{\Sigma}_\mathbf{A} \in \mathbb{R}^{r\times r}$ *for* $r = \mathrm{rank}(\mathbf{A})$ *as in Definition 2.22. Then, for any generalized inverse* $\mathbf{A}^g$*, as in Definition 2.25, there exists matrices* $\mathbf{X} \in \mathbb{R}^{r\times(n-r)}, \mathbf{Y} \in \mathbb{R}^{(p-r)\times r}, \mathbf{Z} \in \mathbb{R}^{(p-r)\times(n-r)}$ *such that*

$$\mathbf{A}^g = \mathbf{V}_\mathbf{A} \begin{bmatrix} \mathbf{\Sigma}_\mathbf{A}^{-1} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix} \mathbf{U}_\mathbf{A}^\mathsf{T}. \tag{2.37}$$

In particular, the Moore–Penrose pseudoinverse $\mathbf{A}^+$ *corresponds to the case* $\mathbf{X} = \mathbf{Y} = \mathbf{Z} = \mathbf{0}$*. In addition, we have that:*

1. *if* $\mathbf{A}$ *has full row rank (implying* $p \geq n$ *and* $\mathbf{A}^\mathsf{T}\mathbf{A}$ *non-singular), then* $\mathbf{A}^+ = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}$*; and*

2. *if* $\mathbf{A}$ *has full column rank (implying* $p \leq n$ *and* $\mathbf{A}\mathbf{A}^\mathsf{T}$ *non-singular), then* $\mathbf{A}^+ = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T})^{-1}$.

Remark 2.28 (**Minimum norm solution with Moore–Penrose pseudoinverse**)**.** It follows from Theorem 2.26 by taking $\mathbf{w} = \mathbf{0}$ that, if the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ admits a solution, then the minimum (Euclidean) norm solution is given by Moore–Penrose pseudoinverse $\mathbf{x} = \mathbf{A}^+\mathbf{b}$. That is, the solution $\hat{\mathbf{x}} = \mathbf{A}^+\mathbf{b}$ is the minimum solution to Equation (2.33):

$$\arg\min_{\mathbf{A}\mathbf{x}=\mathbf{b}} \|\mathbf{x}\|_2 = \mathbf{A}^+\mathbf{b}. \tag{2.38}$$

In case where $\mathbf{b}$ does *not* belong to the column space of $\mathbf{A}$, the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ does *not* admit a solution. In that case though, we can discuss the "closest" solution $\mathbf{x}$ so that the linear system of equation holds approximately $\mathbf{A}\mathbf{x} \approx \mathbf{b}$. When using the Euclidean norm distance to measure this "closeness" of solution, this is the *least squares* solution.

Definition 2.29 (**Least squares and ridge regression**)**.** *For* $\mathbf{A} \in \mathbb{R}^{p\times n}$ *and* $\mathbf{b} \in \mathbb{R}^p$*, the* least squares solution $\mathbf{x}_{\mathrm{LS}} \in \mathbb{R}^n$ *to the linear system in Definition 2.24 is given by*

$$\mathbf{x}_{\mathrm{LS}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2. \tag{2.39}$$

As we shall see below in Theorem 2.30, this in fact defines a set $\mathcal{X}_{\mathrm{LS}}$ *of feasible solutions. We can similarly define the* ridge regression solution $\mathbf{x}_\gamma \in \mathbb{R}^n$ *to the linear system in Definition 2.24 as*

$$\mathbf{x}_\gamma = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \gamma\|\mathbf{x}\|_2^2\right), \tag{2.40}$$

for some $\gamma > 0$ *that penalizes the Euclidean norm of the solution.*

Note that the least squares solution in Definition 2.29 is of particular interest when $\mathbf{b}$ does *not* belong to the column space of $\mathbf{A}$ so that $\mathbf{A}\mathbf{x} = \mathbf{b}$ does *not* admit a solution. Otherwise, $\mathbf{x}_{\text{LS}}$ is just one of the solutions to the linear system given by the generalized (e.g., Moore–Penrose) inverse as in Theorem 2.26.

It turns out that the least squares solution in Definition 2.29 may not be unique and is characterized in the following result.

Theorem 2.30 (**Characterization of least squares and ridge regression solutions**). *For $\mathbf{A} \in \mathbb{R}^{p\times n}$ and $\mathbf{b} \in \mathbb{R}^p$, the least squares solution $\mathbf{x}_{\text{LS}}$ in Definition 2.29 always exists, and all solution are given by*

$$\mathbf{x}_{\text{LS}}(\mathbf{w}) = \mathbf{A}^+\mathbf{b} + (\mathbf{I}_n - \mathbf{A}^+\mathbf{A})\mathbf{w}, \tag{2.41}$$

for arbitrary *vector $\mathbf{w} \in \mathbb{R}^n$ and $\mathbf{A}^+$ the Moore–Penrose pseudoinverse of $\mathbf{A}$. On the other hand, the ridge regression solution, for any $\gamma > 0$ exists and is uniquely (and equivalently) given by*

$$\mathbf{x}_\gamma = (\mathbf{A}^\mathsf{T}\mathbf{A} + \gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}\mathbf{b} = \mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \gamma\mathbf{I}_p)^{-1}\mathbf{b}. \tag{2.42}$$

Remark 2.31 (**Minimum norm least squares solution with Moore–Penrose pseudoinverse**). It follows from Theorem 2.30 by taking $\mathbf{w} = \mathbf{0}$ that, the minimum (Euclidean) norm least squares solution is given by Moore–Penrose pseudoinverse $\mathbf{x}_{\text{LS}}(\mathbf{w} = \mathbf{0}) = \mathbf{A}^+\mathbf{b}$. That is,

$$\arg\min_{\mathbf{x}\in\mathcal{X}_{\text{LS}}} \|\mathbf{x}\|_2 = \mathbf{A}^+\mathbf{b}, \tag{2.43}$$

where $\mathcal{X}_{\text{LS}}$ is the set of feasible least square solutions as in Definition 2.29. Moreover, it follows from the SVD of $\mathbf{A}$ and Theorem 2.30 that

$$\lim_{\gamma\downarrow 0} \mathbf{x}_\gamma = \mathbf{A}^+\mathbf{b}. \tag{2.44}$$

That is, the Moore–Penrose pseudoinverse solution $\mathbf{x}_{\text{LS}}(\mathbf{w} = \mathbf{0}) = \mathbf{A}^+\mathbf{b}$ also corresponds to the "ridgeless" regression solution as $\gamma \to 0$.

Despite arising in many scenarios when, e.g., considering the minimum norm solution to linear system or least squares in Remark 2.28 and 2.31, respectively, the Moore–Penrose pseudoinverse solution $\mathbf{A}^+\mathbf{b}$ can be *numerically unstable*, in that it does *not* depend on $\mathbf{A}$ in a continuous fashion, as opposed to the ridge regularized inverse $(\mathbf{A}^\mathsf{T}\mathbf{A} + \gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}$ or $\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \gamma\mathbf{I}_p)^{-1}$. This is discussed in the following remark.

Remark 2.32 (**Discontinuity of pseudoinverse**). The Moore–Penrose pseudoinverse $\mathbf{A}^+$ of $\mathbf{A}$ maps a (small) singular value $\sigma_i(\mathbf{A})$ to $1/\sigma_i(\mathbf{A})$ and does not depend continuously on $\mathbf{A}$. On the other hand, the regularized inverse $(\mathbf{A}^\mathsf{T}\mathbf{A} + \gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}$ or $\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T} + \gamma\mathbf{I}_p)^{-1}$ maps a (small) singular value $\sigma_i(\mathbf{A})$ to $\frac{\sigma_i(\mathbf{A})}{\gamma+\sigma_i^2(\mathbf{A})}$ and depends on $\mathbf{A}$ in a more "continuous" fashion, but shrinks to the Moore–Penrose pseudoinverse as $\gamma \to 0$. As an example, consider a small rank-one perturbation $\mathbf{A}_\varepsilon$ of a given matrix $\mathbf{A}$ having rank r with SVD $\mathbf{A} = \sum_{i=1}^r \sigma_i(\mathbf{A})\mathbf{u}_i\mathbf{v}_i^\mathsf{T}$ given by (the correspond SVD as)

$$\mathbf{A}_\varepsilon = \mathbf{A} + \varepsilon\mathbf{u}_{r+1}\mathbf{v}_{r+1}^\mathsf{T}, \tag{2.45}$$

for some small $\varepsilon > 0$. Then, by Theorem 2.27, its pseudoinverse $\mathbf{A}_\varepsilon^+$ is given by

$$\mathbf{A}_\varepsilon^+ = \mathbf{A}^+ + \frac{1}{\varepsilon}\mathbf{v}_{r+1}\mathbf{u}_{r+1}^\mathsf{T}, \tag{2.46}$$

and therefore

$$\frac{\|\mathbf{A}_\varepsilon^+ - \mathbf{A}^+\|}{\|\mathbf{A}^+\|} = \frac{\sigma_r(\mathbf{A})}{\varepsilon} \gg 1, \tag{2.47}$$

for ε small. On the other hand, ridge regularized inverse $(\mathbf{A}^\mathsf{T}\mathbf{A}+\gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}$ or $\mathbf{A}^\mathsf{T}(\mathbf{A}\mathbf{A}^\mathsf{T}+\gamma\mathbf{I}_p)^{-1}$ is a more "continuous" function of $\mathbf{A}$ since

$$(\mathbf{A}^\mathsf{T}\mathbf{A}+\gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T} = \mathbf{V}_\mathbf{A}^\mathsf{T}\begin{bmatrix}(\mathbf{\Sigma}_\mathbf{A}^2+\gamma\mathbf{I}_r)^{-1}\mathbf{\Sigma}_\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{U}_\mathbf{A}^\mathsf{T}, \tag{2.48}$$

so that

$$\frac{\|(\mathbf{A}_\varepsilon^\mathsf{T}\mathbf{A}_\varepsilon+\gamma\mathbf{I}_n)^{-1}\mathbf{A}_\varepsilon^\mathsf{T}-(\mathbf{A}^\mathsf{T}\mathbf{A}+\gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}\|}{\|(\mathbf{A}^\mathsf{T}\mathbf{A}+\gamma\mathbf{I}_n)^{-1}\mathbf{A}^\mathsf{T}\|} = \frac{2\sigma_i(\mathbf{A})}{\sigma_i(\mathbf{A})^2/\varepsilon+\varepsilon} \leq 1. \tag{2.49}$$

if we take $\gamma = \sigma_i(\mathbf{A})$ for some $i \in \{1,\ldots,\min(n,p)\}$.

Chapter 3

Linearizing high-dimensional nonlinear functions

There are two motivations for the techniques described in this chapter.

1. First, many ML models are *nonlinear*. For instance, kernel methods extract *nonlinear* features of input data by "lifting" them into some (typically infinitely dimensional) reproducing kernel Hilbert space [27]; and neural networks perform nonlinear classification or regression of input data by using *nonlinear* activation functions [16]. See **??** for more detailed treatments of these nonlinear ML models.

2. Second, linear analysis tools (e.g., basic single-variable calculus, *linear* algebra, random *matrix* theory, etc.) are so powerful that when we encounter nonlinear problems, a common strategy is to find and solve a related approximate linear problem.

This second motivation, of course, holds throughout applied mathematics, science, and engineering; but many of the issues that arise in modern ML mean that we need to revisit these ideas in a broader context. The standard example of this *linearization* approach is provided by the Taylor expansion in calculus: given a deterministic single-variable function $f : \mathbb{R} \to \mathbb{R}$, we can approximate its behavior at a point x near a reference point τ as

$$\begin{aligned} f(x) &= f(\tau) + f'(\tau)(x-\tau) + \frac{f''(\tau)}{2}(x-\tau)^2 + \dots \\ &\approx f(\tau) + f'(\tau)(x-\tau), \end{aligned}$$

where the approximation ($\approx$) in the second line holds when the function f is sufficiently smooth so that the remaining higher-order terms are small and can be ignored. When this approximation holds, the function $f(\cdot)$ is well-approximated by a linear/affine function.

In this chapter, we are interested in the generalization of these "linearization" ideas from single-variable deterministic functions to high-dimensional random functions of the form $f(x) : \mathbb{R}^n \to \mathbb{R}$. In ML, the variable x is typically a high-dimensional vector, $\mathbf{x} \in \mathbb{R}^n$, in which case $f(\mathbf{x}) \in \mathbb{R}$ may be interpreted as a "scalar observation" of that random vector (as in Definition 1.17 of Chapter 1.4). We will discuss different approaches to assess the behavior of the nonlinear function $f(\mathbf{x})$ (or its statistics such as the expectation $\mathbb{E}[f(\mathbf{x})]$), depending on the properties of $f(\cdot)$, the random $\mathbf{x}$, and the dimension n. To accomplish this, we need to perform some sort of *high-dimensional linearization*. In Chapter 3.1, we will present two different scaling regimes that are particularly relevant for modern ML. In Chapter 3.2, we will describe how the Taylor expansion approach can be applied not just to single variable deterministic functions but also to certain high-dimensional random functions in one of these scaling regimes. In Chapter 3.3, we will describe how a more sophisticated but complementary linearization approach can be applied in the other scaling regime. Finally, in Chapter 3.4, we

will introduce the idea of Linear Equivalent that unifies both approaches and propose High-dimensional Equivalents (as in Definition 1.1) by linearizing nonlinear functions.

3.1 Two different scaling regimes of $f(\mathbf{x})$

We start by recalling the two scaling regimes (the LLN regime and the CLT regime) that we have reviewed in Chapter 1.2, under the form of generic scalar observations of large-dimensional random vectors.

Two scaling regimes

Definition 3.1 (**Two scaling regimes**). *For a scalar observation $f(\mathbf{x})$ of a large-dimensional random vector $\mathbf{x} \in \mathbb{R}^n$ via some $f\colon \mathbb{R}^n \to \mathbb{R}$, consider the following two* scaling regimes*:*

1. ***LLN regime**: this holds when $f(\mathbf{x})$ establishes, for n large, a* LLN-type *concentration, strongly concentrating around a deterministic quantity, say $\mathbb{E}[f(\mathbf{x})]$, in such a way that its distribution function becomes (asymptotically)* degenerate*, e.g., $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})] \to 0$ in probability or almost surely as $n \to \infty$.*

2. ***CLT regime**: this holds when $f(\mathbf{x})$ establishes, for n large, a* CLT-type *concentration, remaining random, and having a* non-degenerate *distribution function in the $n \to \infty$ limit, e.g., $\sqrt{n}\,(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]) \to \mathcal{N}(0,1)$ in distribution as $n \to \infty$.*

In the following example, we describe how different objects from Chapter 2 (norms, inner products, and angles) behave in two different scaling regimes (of the value of the dimension n) in Definition 3.1.

Example 3.2 (**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 (the scaling by $\sqrt{n}$ is made so that $\mathbb{E}[\|\mathbf{x}\|_2^2] = 1$, as in Remark 2.4), and $\mathbf{y} \in \mathbb{R}^n$ be a* deterministic *vector of unit norm $\|\mathbf{y}\| = 1$; and consider the following* nonlinear *objects of interest with a nonlinear function $\phi\colon \mathbb{R} \to \mathbb{R}$ acting in two different regimes:*

1. ***LLN regime**: here, we consider random variables*

$$f_{\mathrm{LLN}}(\mathbf{x}) = \|\mathbf{x}\|^2 \quad \text{or} \quad f_{\mathrm{LLN}}(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{y}, \tag{3.1}$$

   that establish a LLN-type concentration, as $n \to \infty$, and we are interested in the nonlinear $\phi(f_{\mathrm{LLN}}(\mathbf{x}))$; and

2. ***CLT regime**: here, we consider random variables*

$$f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|^2 - 1) \quad \text{or} \quad f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^{\mathsf{T}}\mathbf{y}, \tag{3.2}$$

   that establish a CLT-type concentration, as $n \to \infty$, and we are interested in the nonlinear $\phi(f_{\mathrm{CLT}}(\mathbf{x}))$.

The two regimes in Example 3.2 follow from the two well-known convergence results (recall from Remark 2.4 on the difference scaling for inner products and norms):

1. the (strong) **law of large numbers (LLN)** in Theorem 1.7, which implies that

$$\|\mathbf{x}\|^2 \to \mathbb{E}[\mathbf{x}^{\mathsf{T}}\mathbf{x}] = 1 \quad \text{and} \quad \mathbf{x}^{\mathsf{T}}\mathbf{y} \to \mathbb{E}[\mathbf{x}^{\mathsf{T}}\mathbf{y}] = 0, \tag{3.3}$$

   almost surely as $n \to \infty$; and

2. the **central limit theorem (CLT)** in Theorem 1.8, which implies that

$$\sqrt{n}(\|\mathbf{x}\|^2 - 1) \to \mathcal{N}(0, 2) \quad \text{and} \quad \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \to \mathcal{N}(0, 1), \tag{3.4}$$

in law as $n \to \infty$.

These two results can be written, as in Remark 2.4, in the following more compact form:

$$\|\mathbf{x}\|^2 \simeq 1 + \mathcal{N}(0, 2)/\sqrt{n} \quad \text{and} \quad \mathbf{x}^\mathsf{T}\mathbf{y} \simeq 0 + \mathcal{N}(0, 1)/\sqrt{n}, \tag{3.5}$$

for n large.

Remark 3.3 (**Different possible scalings for the random variable**)**.** The two scaling regimes (of scalar observations of large random vectors) defined in Definition 3.1 are of particular interest when being evaluated though some nonlinear function $\phi\colon \mathbb{R} \to \mathbb{R}$ as in Example 3.2. As we shall see below in Chapter 3.2 and Chapter 3.3, the behavior of such nonlinear random variables depends on different properties of ϕ in different scaling regimes.

Note also that beyond the LLN and CLT, there are other (but trivial) scaling regimes: Consider a random vector $\mathbf{x} \in \mathbb{R}^n$ having zero mean and unit variance entries (so that $\|\mathbf{x}\|^2 \to n$), a nonlinear function $\phi\colon \mathbb{R} \to \mathbb{R}$ could act on the following scaling regimes:

1. **LLN regime**: $f_{\text{LLN}}(\mathbf{x}) = \|\mathbf{x}\|^2/n$ or $f_{\text{LLN}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}//\sqrt{n}$; and
2. **CLT regime**: $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|^2/n - 1)$ or $f_{\text{CLT}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}$; and
3. **trivial regimes** at ∞: $f(\mathbf{x}) = \|\mathbf{x}\|^2/C_n$ for any $C_n = o(n)$, for which we have $f(\mathbf{x}) = \|\mathbf{x}\|^2/C_n \to \infty$ as $n \to \infty$; and similarly $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}/C_n$ for any $C_n = o(1)$, for which we have $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}/C_n \to \infty$ as $n \to \infty$.

Remark 3.4 (**LLN- versus CLT-type concentration**)**.** Here and in the following, we refer to the LLN-type results (in the first item of Example 3.2) as *LLN-type concentration*, since random variables of the form $f_{\text{LLN}}(\mathbf{x}) = \|\mathbf{x}\|^2$ or $\mathbf{x}^\mathsf{T}\mathbf{y}$ are close-to-deterministic and exhibit deterministic-like or *degenerate* behavior for n large. Similarly, we refer to the CLT-type results (in the second item of Example 3.2) as *CLT-type concentration*, since random variables of the form $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|^2 - 1)$ or $\sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y}$ are *not* close-to-deterministic; instead, they remain inherently random and exhibit a *non-degenerate* distribution function for n large.

It is worth clarifying that these two categories of concentration—LLN-type and CLT-type—are subfields of high-dimensional *concentration* results in the literature of high-dimensional probability and statistics [19, 36, 38]. These results provide a framework to describe, e.g., the sub-gaussian tail behavior of random variable around (or away from) from their expectations.

The significance of this discussion is that the "scalings" of the two families of nonlinear objects in Example 3.2 are different (and thus their linearizations will need to be different).

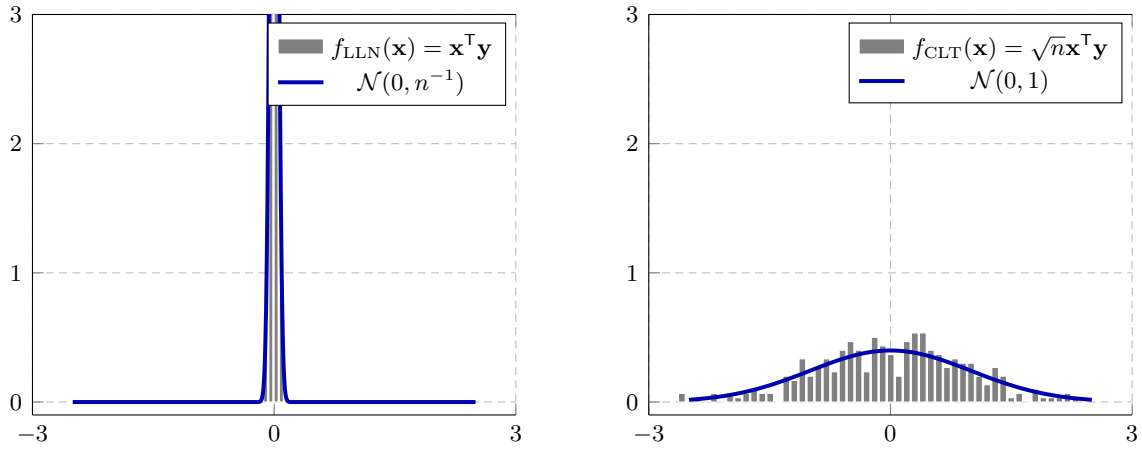
1. **LLN regime.** For objects in the LLN regime, the nonlinear function ϕ is applied on a *close-to-deterministic* quantity, in the sense that

$$\|\mathbf{x}\|^2 = 1 + O(n^{-1/2}) \text{ and } \mathbf{x}^\mathsf{T}\mathbf{y} = 0 + O(n^{-1/2}), \tag{3.6}$$

with high probability for n large, due to the dominant LLN behavior. In this case, the familiar Taylor expansion approach (from deterministic single-variable calculus) will suffice, even if the justification is slightly different since $\mathbf{x}$ is a random variable.

2. **CLT regime.** For objects in the CLT regime, the nonlinear function ϕ is applied on a normally distributed *random* variable. As a consequence of the CLT, that is *not* close to a deterministic quantity, in the sense that for n large,

$$\sqrt{n}(\|\mathbf{x}\|^2 - 1) \sim \mathcal{N}(0, 2) \text{ and } \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0, 1), \tag{3.7}$$



(a) LLN regime (b) CLT regime

Figure 3.1: Illustrations of the random variables $\mathbf{x}^\mathsf{T}\mathbf{y}$ in the LLN (Figure 3.1a) and the CLT (Figure 3.1b) regime, with $n = 500$. In the LLM regime, the random variable concentrates strongly around its expected value, with very small variability; while in the CLT regime, the random variable still have substantial variability about its expected value.

and are in particular *not* close the mean zero. In this case, more sophisticated high-dimensional approaches based on orthogonal polynomials will be needed to perform the linearization.

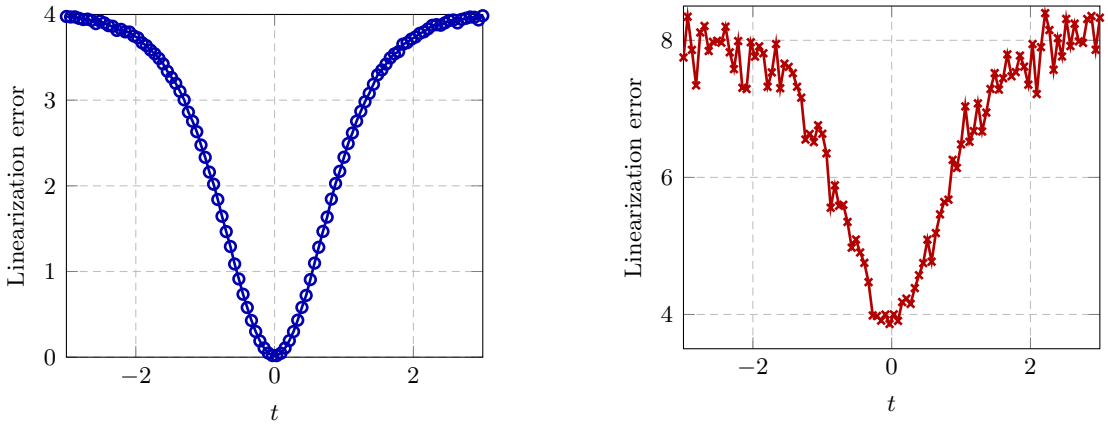
Figure 3.1 visualizes the behavior of inner-products $\mathbf{x}^\mathsf{T}\mathbf{y}$ of Example 3.2 in the LLN regime (where $f_{\mathrm{LLN}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y} \simeq \mathcal{N}(0, n^{-1})$ is almost a Dirac delta function at zero for n large) and the CLT regime (where $f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n}\mathbf{x}^\mathsf{T}\mathbf{y} \simeq \mathcal{N}(0, 1)$ "spreads" out on the axis, on a scale that is comparable to the range over which a "quadratic approximation" using Taylor expansion to $\phi(\cdot)$ would be valid). Figure 3.2 compares the linearization (mean squared) errors $(\phi(f(\mathbf{x})) - \phi(t))^2$ by considering all possible deterministic $t \in [-3, 3]$, for random variables $f_{\mathrm{LLN}}(\mathbf{x})$ and $f_{\mathrm{CLT}}(\mathbf{x})$ in the LLN and CLT regime, respectively. We observe that by going over all possible deterministic value of $\phi(t), t \in [-3, 3]$, the linearization error of $\phi(f_{\mathrm{LLN}}(\mathbf{x}))$ in the LLN regime can be reduced to zero, but this is not the case for $\phi(f_{\mathrm{CLT}}(\mathbf{x}))$ in the CLT regime. In particularly, the linearization errors for $\phi(f_{\mathrm{CLT}}(\mathbf{x}))$ using any $\phi(t), t \in [-3, 3]$ remains rather random, and is empirically observed constantly larger than 4.

These two different linearization approaches—via the *Taylor expansion* and via *orthogonal polynomials*—are summarized in Table 3.1. They are discussed in Chapter 3.2 and Chapter 3.3 below, respectively. In particular, we will better understand the observation in Figure 3.2 that a small linearization error can be achieved by approximating the nonlinear $\phi(f(\mathbf{x}))$ in a deterministic fashion (i.e., by $\phi(t)$), but *only* in the LLN regime, *not* in the CLT regime.

3.2 Linearization via Taylor expansion

In this section, we will describe the Taylor expansion approach for linearizing nonlinear functions. Although most well-known for being applied to deterministic single-variable functions, the method also applies to certain high-dimensional random functions (basically, those in the LLM regime).

Taylor expansion is perhaps the most popular approach to perform *local* linearization of a *smooth* nonlinear function. Here is the basic result for real-valued functions of a single variable.



(a) LLN regime

(b) CLT regime

Figure 3.2: Illustrations of the (mean squared) linearization errors $(\phi(f(\mathbf{x})) - \phi(t))^2$ for $\phi(t) = \tanh(t)$ as in Example 3.14, by searching for all possible $t \in [-3, 3]$, in the LLN regime (Figure 3.2a) with $f_{\text{LLN}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}$ and in the CLT regime (Figure 3.2b) with $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n}\mathbf{x}^\mathsf{T}\mathbf{y}$, for $n = 256$, and errors are obtained over 1 024 samples. We observe that while there exist deterministic $t \in [-3, 3]$ (in fact around $t = 0$) such that the linearization error of $\phi(\cdot)$ can be made small (i.e., close to zero) in the LLN regime, this is not the case in the CLT regime. In the CLT regime, the linearization error of $\phi(f_{\text{CLT}}(\mathbf{x}))$ is always larger than 4, for any $t \in [-3, 3]$.

Theorem 3.5 (Taylor's theorem for deterministic single-variable functions, [26, Theorem 8.4]). *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ 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*

$$\phi(x) = \phi(\tau) + \phi'(\tau)(x-\tau) + \frac{\phi''(\tau)}{2}(x-\tau)^2 + \ldots + \frac{\phi^{(k)}(\tau)}{k!}(x-\tau)^k + h_k(x)(x-\tau)^k, \quad (3.8)$$

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

In particular, for a deterministic single variable x, Theorem 3.5 applies to assess the local behavior of $\phi(x)$ around $x = \tau + o(1)$ as a low-degree polynomial that contains both linear (i.e., $\phi'(\tau)(x-\tau)$) and nonlinear (e.g., quadratic or higher-order) components, in the sense that

$$\phi(x) = \phi(\tau) + \phi'(x-\tau) + \frac{\phi''(\tau)}{2}(x-\tau)^2 + o(x-\tau)^2. \quad (3.9)$$

In the following, we discuss how the familiar Taylor expansion approach in Theorem 3.5 can be applied to linearize certain nonlinear functions ϕ of interest, as in Example 3.2. In particular, Theorem 3.5 can be applied in an operational sense to the LLN regime.

What makes the Taylor expansion approach in Theorem 3.5 work? To apply the Taylor expansion approach in Theorem 3.5 to linearize a nonlinear transformation $\phi(x)$ of the (deterministic or random) variable x, the main technical requirements are the following.

1. **Smoothness.** The nonlinear function ϕ under study should be *smooth* (or, more properly speaking, continuously differentiable), at least in the neighborhood of the point τ of interest, so that the derivatives $\phi'(\tau), \phi''(\tau), \ldots$ make sense.

2. **LLN-type concentration.** The variable of interest x is sufficiently close to (or, concentrates around, when being random) the point τ so that the higher orders terms are

Table 3.1: Two different scaling regimes and their corresponding high-dimensional linearization approaches.

Scaling regime	LLN regime	CLT regime
Linearization technique	Taylor expansion in Theorem 3.5 of Chapter 3.2	Orthogonal polynomial in Theorem 3.10 of Chapter 3.3
Smoothness of ϕ	Locally smooth ϕ	Possibly non-smooth ϕ
Object of interest $\phi(f(\mathbf{x}))$ for $\phi\colon \mathbb{R}\to\mathbb{R}$, $f\colon \mathbb{R}^n\to\mathbb{R}$ in Example 3.2	$f_{\mathrm{LLN}}(\mathbf{x}) = \|\mathbf{x}\|^2$ or $\mathbf{x}^\mathsf{T}\mathbf{y}$ in Equation (3.1)	$f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|^2-1)$ or $\sqrt{n}\cdot\mathbf{x}^\mathsf{T}\mathbf{y}$ in Equation (3.2)
Linearization result	$\phi(f_{\mathrm{LLN}}(\mathbf{x}))$ in Proposition 3.6	$\mathbb{E}[\phi(f_{\mathrm{CLT}}(\mathbf{x}))]$ in Proposition 3.12
Object of interest $f(\phi(\cdot))$ for entry-wise $\phi\colon \mathbb{R}^p\to\mathbb{R}^p$ $f\colon \mathbb{R}^p\to\mathbb{R}$ in Example 3.16	$\phi_{\mathrm{LLN}}(\mathbf{X}\mathbf{y})$, $\mathbf{X}\in\mathbb{R}^{p\times n}, \mathbf{y}\in\mathbb{R}^n$ via observation $f\colon \mathbb{R}^p\to\mathbb{R}$	$\phi_{\mathrm{CLT}}(\sqrt{n}\cdot\mathbf{X}\mathbf{y})$, $\mathbf{X}\in\mathbb{R}^{p\times n}, \mathbf{y}\in\mathbb{R}^n$ via observation $f\colon \mathbb{R}^p\to\mathbb{R}$
Linearization result	$f(\phi_{\mathrm{LLN}}(\mathbf{X}\mathbf{y}))$ in Proposition 3.18 for $f(\cdot)=\mathbf{a}^\mathsf{T}(\cdot)/\sqrt{p}$	$f(\phi_{\mathrm{CLT}}(\sqrt{n}\cdot\mathbf{X}\mathbf{y}))$ in Proposition 3.19 for $f(\cdot)=\mathbf{a}^\mathsf{T}(\cdot)/\sqrt{p}$

neglectable (or, more properly speaking, so that the Taylor series is convergent). A more detailed discussion of these two points is provided below.

Extending Taylor's theorem to high-dimensional random functions. To use Taylor's theorem in Theorem 3.5 to assess the nonlinear behavior of $\phi(x)$ for some random variable x, e.g., as those in Example 3.2, it suffices to show that order control (i.e., the $o(x-\tau^2)$ term in Equation (3.9)) holds with some (high) probability. Here is the basic result for the two families of nonlinear examples from Example 3.2, in the LLN regime.

Proposition 3.6 (**Taylor expansion of high-dimensional random functions in the LLN regime**). *For random variable $f_{\mathrm{LLN}}(\mathbf{x}) = \|\mathbf{x}\|^2$, with $\sqrt{n}\mathbf{x}\in\mathbb{R}^n$ having i.i.d. standard Gaussian entries, in the LLN regime (as in the first item of Example 3.2), it follows from the LLN that $\|\mathbf{x}\|^2-1\to 0$, and from the CLT that $\|\mathbf{x}\|^2-1 = O(n^{-1/2})$, with high probability for n large. Thus, it follows from Theorem 3.5 and the differentiability of ϕ that*

$$\phi(\|\mathbf{x}\|^2) = \phi(1) + \phi'(1)\underbrace{(\|\mathbf{x}\|^2-1)}_{O(n^{-1/2})} + \frac{1}{2}\phi''(1)\underbrace{(\|\mathbf{x}\|^2-1)^2}_{O(n^{-1})} + O(n^{-3/2}), \tag{3.10}$$

with high probability. Similarly, for random variable $f_{\mathrm{LLN}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}$ with $\|\mathbf{y}\|=1$, it follows that

$$\phi(\mathbf{x}^\mathsf{T}\mathbf{y}) = \phi(0) + \phi'(0)\underbrace{\mathbf{x}^\mathsf{T}\mathbf{y}}_{O(n^{-1/2})} + \frac{1}{2}\phi''(0)\underbrace{(\mathbf{x}^\mathsf{T}\mathbf{y})^2}_{O(n^{-1})} + O(n^{-3/2}), \tag{3.11}$$

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

Remark 3.7 (**Delta method**). By ignoring second and higher-order terms in Proposition 3.6, the results in Equation (3.10) and Equation (3.11) can be rewritten as

$$\sqrt{n}\left(\phi(\|\mathbf{x}\|^2) - f(1)\right) \xrightarrow{d} \mathcal{N}\left(0, 2(\phi'(\tau))^2\right),$$
$$\sqrt{n}\left(\phi(\mathbf{x}^\mathsf{T}\mathbf{y}) - f(0)\right) \xrightarrow{d} \mathcal{N}\left(0, (\phi'(\tau))^2\right).$$

This is known in the literature as the Delta method; see, e.g., [35, Chapter 3].

In the following, we discuss in more detail the two working assumptions of Theorem 3.5.

Smoothness assumptions. Regarding the smoothness assumption, one can relax it. For example, for a *non-smooth* and nonlinear function ϕ, one can evaluate the *expected* behavior $\mathbb{E}[\phi(x)]$ of $\phi(x)$, for x being random. While the function ϕ may not be differentiable everywhere (and in particular, in the neighborhood $x = \tau$ of interest), it can still have almost everywhere weak derivative ϕ' (in the sense of distributions, see for example [31, Section 3] for an introduction) such that

$$\int \phi'(t)\mu(dt) = \mathbb{E}[\phi'(x)] < \infty, \tag{3.12}$$

exists, for random variable x having law μ. In a sense, for non-differential ϕ, ϕ' does not exists in the sense of ordinary functions, but we can still define such derivative of ϕ in a weak sense, so long that the integral $\int \phi'(t)\mu(dt)$ exists for some (signed) Borel measure μ.

A concrete example of this in the case of a standard Gaussian x is known as Stein's lemma, which states: For standard Gaussian random variable $x \sim \mathcal{N}(0, 1)$, we have that

$$\mathbb{E}[\phi'(x)] = \mathbb{E}[x\phi(x)], \tag{3.13}$$

as long as the right-hand-side term is finite. The proof of this result follows from the integration by parts formula as

$$\mathbb{E}[x\phi(x)] = \int t\phi(t)\mu(dt) = \int \phi(t)\frac{1}{\sqrt{2\pi}}te^{-\frac{t^2}{2}}dt = \int \phi'(t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt = \mathbb{E}[\phi'(x)], \tag{3.14}$$

with standard Gaussian measure $\mu(dt) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$. This result allows one to assess the expectation $\mathbb{E}[\phi'(x)]$ for standard Gaussian x and weakly differentiable ϕ.

LLT-type concentration assumption. The LLN-type concentration assumption is a more intrinsic limitation of the Taylor expansion approach, as this approach allows one to assess *only* the *local* behavior of the nonlinear function $\phi(x)$ around some $x = \tau$.

A concrete example of this arises in the proof of Proposition 3.6,[6] which strongly relies on the fact that both $\|\mathbf{x}\|^2 - 1$ and $\mathbf{x}^\mathsf{T}\mathbf{y}$ are of order $O(n^{-1/2})$ with high probability, which happens in the LLN regime. Otherwise, the higher-orders terms in Theorem 3.5 cannot be ignored (at least with high probability). In particular, in the CLT regime, the nonlinear function ϕ is applied on Gaussian random variables that do *not* exhibit this type of strong concentration around any deterministic quantity (in the sense that the random fluctuation vanishes, e.g., as the dimension n grows). In this setting, it no longer makes sense to apply the Taylor expansion approach in Theorem 3.5, since the higher-order terms cannot be ignored.

3.3 Linearization via orthogonal polynomial expansion

In this section, we will discuss a different linearization method, the *orthogonal polynomial approach*, which can be applied to high-dimensional random functions, in particular those in the CLT regime. Among other things, this approach allows one to characterize the behavior of the nonlinear function $\mathbb{E}[\phi(x)]$ of *random variable* x that, in particular, does *not* strongly concentrate around a point of interest τ, and instead exhibits a CLT-type concentration. These functions cannot be linearized using Taylor expansion technique in Theorem 3.5, due to their "non-LLN-type concentration" and the "non-smooth" properties of such x.

To understand the orthogonal polynomial approach, we can take, for random x, a functional analysis perspective[7] on the expectation $\mathbb{E}[\phi(x)]$. This is different from the Taylor expansion

[6] See [35, Chapter 2] for a detailed proof.

[7] That is, we use ideas from deterministic functional analysis to assess and explain the expected behavior of nonlinear random variables (e.g., in the CLT regime of Example 3.2). This should be compared and contrasted to the use of *deterministic* Taylor expansion to treat *random* but close-to-deterministic nonlinear random variables (e.g., in the LLN regime of Example 3.2).

perspective that we saw in Chapter 3.2 that viewed ϕ as a mapping from $\mathbb{R} \to \mathbb{R}$,

A functional analysis perspective of $\mathbb{E}[\phi(x)]$. Consider the following *functional analysis perspective* of the expectation $\mathbb{E}[\phi(x)]$. For a generic random variable x following some law μ, the expectation $\mathbb{E}[\phi(x)]$ of the nonlinear transformation $\phi(x)$ can be expressed as

$$\mathbb{E}_{x\sim\mu}[\phi(x)] = \int \phi(t)\mu(dt). \tag{3.15}$$

This corresponds to the integral of ϕ with respect to the probability measure μ, for some (deterministic) ϕ living in some (possibly infinite-dimensional) function space.

We know that, in the case of Euclidean space (reviewed in Chapter 2, recall Remark 2.2), 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 expanded as

$$\mathbf{x} = \sum_{i=1}^{n} (\mathbf{x}^\mathsf{T}\mathbf{e}_i)\mathbf{e}_i = \sum_{i=1}^{n} x_i \mathbf{e}_i, \tag{3.16}$$

with the inner product (see Definition 2.1) $\mathbf{x}^\mathsf{T}\mathbf{e}_i = x_i$ being equal to the ith coordinate of $\mathbf{x}$. A similar result holds more generally. In particular, for a function f living in some (potentially infinite dimensional) function space, such an f can be expanded into the sum of "orthonormal" basis functions, weighted by the projection of f onto these basis functions.

The concepts of inner products for functions, families of orthonormal functions in some Hilbert space, and the corresponding orthogonal polynomial expansions are made precise in the following definition.

Orthogonal Polynomials and Orthogonal Polynomial Expansion

Definition 3.8 (Orthogonal polynomials and orthogonal polynomial expansion). *For a probability measure μ, define the inner product between two functions ϕ and ψ as*

$$\langle \phi, \psi \rangle_\mu \equiv \int \phi(x)\psi(x)\mu(dx) = \mathbb{E}[\phi(x)\psi(x)], \tag{3.17}$$

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

$$\langle P_{\ell_1}, P_{\ell_2} \rangle = \mathbb{E}[P_{\ell_1}(x)P_{\ell_2}(x)] = \delta_{\ell_1=\ell_2}. \tag{3.18}$$

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

$$\phi(x) \sim \sum_{\ell=0}^{\infty} a_{\phi;\ell} P_\ell(x), \quad a_{\phi;\ell} = \int \phi(x)P_\ell(x)\mu(dx). \tag{3.19}$$

In Definition 3.8, we used the notation "$\phi \sim \sum_{l=0}^{\infty} a_{\phi;\ell}P_\ell$" to denote that $\|\phi - \sum_{\ell=0}^{L} a_{\phi;\ell}P_\ell\|_\mu \to 0$ as $L \to \infty$ with $\|\phi\|_\mu^2 = \langle \phi, \phi \rangle$, or equivalently

$$\int \left(\phi(t) - \sum_{\ell=0}^{L} a_{\phi;\ell}P_\ell(t)\right)^2 \mu(dt) = \mathbb{E}_{x\sim\mu}\left[\left(\phi(x) - \sum_{\ell=0}^{L} a_{\phi;\ell}P_\ell(x)\right)^2\right] \to 0, \tag{3.20}$$

written in the form of a nonlinear random variable $\phi(x)$. It follows from the Riesz-Fischer theorem (see [26, Theorem 11.43]), that if the family of orthogonal polynomial $\{P_\ell(x)\}_{\ell=0}^{\infty}$ forms a orthonormal basis of $L^2(\mu)$, the set of all square-integrable functions with respect to $\langle \cdot, \cdot \rangle$, then we can expand any ϕ as in Equation (3.20).

Table 3.2: Correspondence between expansions in Hilbert versus Euclidean space.

Type of space:	Euclidean vector space	Hilbert functional space
Definition and notation:	$\mathbb{R}^n$ in Definition 2.1	$L^2(\mu)$ in Definition 3.8
Inner products (and norms):	$\mathbf{x}^\mathsf{T}\mathbf{y} = \sum_{i=1}^n x_i y_i$, $\|\mathbf{x}\|_2^2 = \mathbf{x}^\mathsf{T}\mathbf{x}$	$\langle \phi, \psi \rangle_\mu \equiv \int \phi(x)\psi(x)\mu(dx)$, $\|\phi\|_\mu^2 = \langle \phi, \phi \rangle_\mu$
Expansion:	$\mathbf{x} = \sum_{i=1}^n (\mathbf{x}^\mathsf{T}\mathbf{e}_i)\mathbf{e}_i = \sum_{i=1}^n x_i \mathbf{e}_i$	$\phi(x) \sim \sum_{\ell=0}^\infty a_{\phi;\ell} P_\ell(x)$

Remark 3.9 (**Expansion in Hilbert versus Euclidean space**)**.** We can compare Definition 3.8 for the expansion of functions living in some (infinite-dimensional) Hilbert space, to that for (finite-dimensional) Euclidean vector space in Definition 2.1. We observe the following correspondence:

1. the inner product in Equation (3.17) between functions (measured by μ) extends the inner product in Definition 2.1 between Euclidean vectors;
2. the norm $\|\phi\|_\mu$ of some function extends the Euclidean norm of a vector in Remark 2.2, and both present the total "energy" (of the function ϕ, when measured by the "weight function" μ, and of the finite-dimensional Euclidean vector); and
3. the expansion of functions into in Equation (3.19) extends the canonical basis expansion of Euclidean vectors in Equation (3.16).

See Table 3.2 for an summary of these correspondences. As we shall below, the expansion in Hilbert functional space plays a crucial role in evaluating nonlinear random variables of the form $\phi(f_{\text{CLT}}(\mathbf{x}))$, for $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|^2 - 1)$ or $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y}$ in the CLT regime as in the second item of Example 3.2.

As a specific type of expansions in Hilbert functional space, the orthogonal polynomial expansion given in Equation (3.19) provides the basis for a more sophisticated linearization technique that allows one to assess the behavior of $\mathbb{E}[\phi(x)]$ for *not-close-to-deterministic* scalar random variable x, such as the scalar observation $x = f_{\text{CLT}}(\mathbf{x})$ in the CLT regime. An example of this is provided by $x = f_{\text{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0,1)$ with $\|\mathbf{y}\| = 1$ and $\sqrt{n}\mathbf{x} \in \mathbb{R}^n$ having standard Gaussian entries, where there is non-trivial probability that the Gaussian random variable $x = f_{\text{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y}$ is "spread out" on the real line.

Hermite polynomial expansion. When one is interested in the Gaussian measure, $\mu(dx) = \exp(-x^2/2)/\sqrt{2\pi}$, the natural family of orthogonal polynomials to consider is the normalized *Hermite polynomial family*. Here is the definition.

Theorem 3.10 (**Hermite polynomial expansion, [26, Theorem 11.43]**)**.** *For $x \in \mathbb{R}$, the ℓ^{th} order* normalized *Hermite polynomial, denoted* $\text{He}_\ell(x)$*, is given by given by*

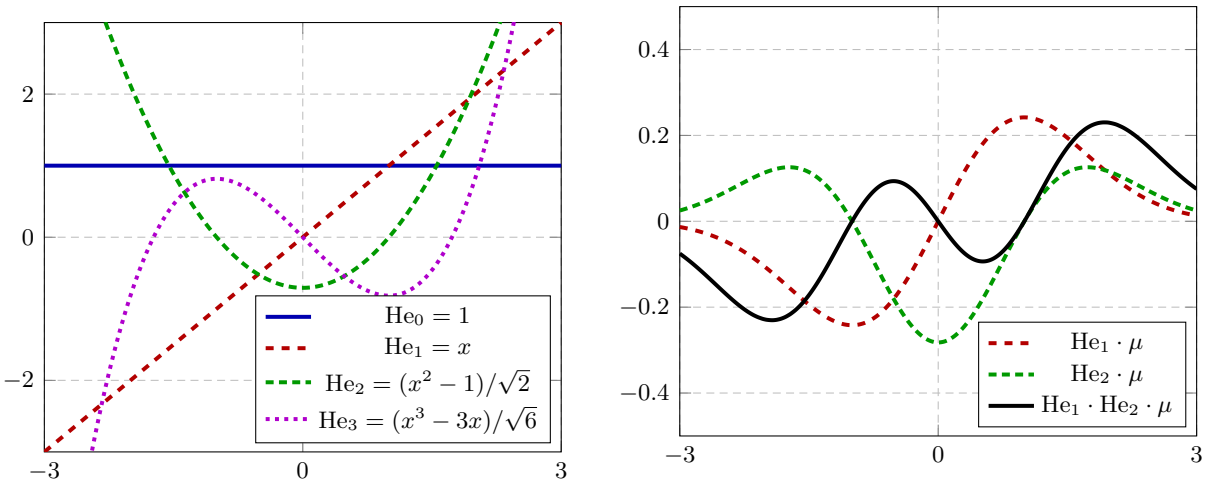
$$\text{He}_0(x) = 1, \text{ and } \text{He}_\ell(x) = \frac{(-1)^\ell}{\sqrt{\ell!}} e^{\frac{x^2}{2}} \frac{d^n}{dx^n}\left(e^{-\frac{x^2}{2}}\right), \text{ for } \ell \geq 1. \tag{3.21}$$

The (normalized) Hermite polynomials

1. *are orthogonal polynomials, and (as the name implies) are orthonormal with respect the standard Gaussian measure, in the sense that*

$$\int \text{He}_m(x)\text{He}_n(x)\mu(dx) = \delta_{nm}, \tag{3.22}$$

for $\mu(dt) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ the standard Gaussian measure;



(a) First four Hermite polynomials.

(b) First- and second-order Hermite polynomial weighted by Gaussian measure.

Figure 3.3: Illustration of the first four Hermite polynomials as in Theorem 3.10 (Figure 3.3a) and of the first- and second-order Hermite polynomial (He_1 and He_2) weighted by the Gaussian measure $\mu(dx) = \exp(-x^2/2)/\sqrt{2\pi}$ (Figure 3.3b).

2. *form an orthonormal basis of $L^2(\mu)$, the Hilbert space consist of all square-integrable functions with respect to the inner product $\langle \phi, \psi \rangle \equiv \int \phi(x)\psi(x)\mu(dx)$; and*

3. *can be used to formally expand any $\phi \in L^2(\mu)$ as*

$$\phi(x) \sim \sum_{\ell=0}^{\infty} a_{\phi;\ell} \mathrm{He}_\ell(x), \quad a_{\phi;\ell} = \int \phi(x)\mathrm{He}_\ell(x)\mu(dx) = \mathbb{E}[\phi(x)\mathrm{He}_\ell(x)], \tag{3.23}$$

where we use '$\phi \sim \sum_{\ell=0}^{\infty} a_{\phi;\ell}\mathrm{He}_\ell$' as in (3.20) of Definition 3.8, for standard Gaussian random variable $x \sim \mathcal{N}(0,1)$. The coefficients $a_{\phi;\ell}$s are generalized *moments of the standard Gaussian measure μ involving ϕ, and we have*

$$a_{\phi;0} = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi(x)], \quad a_{\phi;1} = \mathbb{E}[x\phi(x)], \quad \sqrt{2}a_{\phi;2} = \mathbb{E}[x^2\phi(x)] - a_{\phi;0}, \tag{3.24}$$

as well as

$$\nu_\phi = \mathbb{E}[\phi^2(x)] = \sum_{\ell=0} a_{\phi;\ell}^2. \tag{3.25}$$

As an example of the Hermite polynomials, see Figure 3.3. In Figure 3.3a, we display the first four (normalized) Hermite polynomials. This is in the spirit of the expansion into Fourier basis commonly used in time-frequency analysis (see for example [30]), with the functions now being evaluated with respect to the Gaussian measure μ. In Figure 3.3b, we depict the first- and second-order Hermite polynomial (He_1 and He_2), but weighted by the Gaussian measure $\mu(dx) = \exp(-x^2/2)/\sqrt{2\pi}$. Comparing Figure 3.3b to Figure 3.3a, we see that the two (normalized) Hermite polynomial He_1 and He_2 are indeed "orthogonal" to each other when measured by μ, in the sense that the $\langle \mathrm{He}_1, \mathrm{He}_2 \rangle \equiv \int \mathrm{He}_1(x)\mathrm{He}_2(x)\mu(dx) = 0$.

Remark 3.11 (**Gegenbauer polynomials and beyond**)**.** While Hermite polynomials are probably of greatest interest in ML, we should emphasize how they arise. They arise due to the Gaussian fluctuations in the random variable being linearized. For other ML models with different noise fluctuations, other orthogonal polynomials would be appropriate For example, a different, yet closely related, family of orthogonal polynomial, the Gegenbauer polynomial,

arises naturally in the evaluation of $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ for independent $\mathbf{x}_i, \mathbf{x}_j \sim \text{Unif}\left(\mathbb{S}^{p-1}(\sqrt{p})\right)$ uniformly drawn from the p-dimensional sphere of radius $\sqrt{p}$. See [23] for an application of the family of Gegenbauer polynomial in neural networks, and [14, 32] for more details on general orthogonal polynomials (beyond Hermite and Gegenbauer).

The Hermite polynomial expansion in Theorem 3.10 allows one to approximate, for standard Gaussian random variable $x \sim \mathcal{N}(0,1)$, the *expectation* $\mathbb{E}[\phi(x)]$ of square-integrable nonlinear (and in particular, possibly *non-polynomial*) function $\phi(x)$ using a (sufficiently high-order) *polynomial* function,

$$\phi(x) \sim \tilde{\phi}(x) = \sum_{\ell=0}^{L} a_{\phi;\ell}\text{He}_\ell(x). \tag{3.26}$$

This can be done, in particular, in the CLT regime, where needs to evaluate the *expected* nonlinear behavior of $\phi(\cdot)$ applied on, e.g., the inner-product of the type $\sqrt{n}\cdot\mathbf{x}^\mathsf{T}\mathbf{y}$ that admits an asymptotically Gaussian behavior.

With Theorem 3.10, we get the following linearizations in the CLT regime.

Proposition 3.12 (**Hermite polynomial expansion of high-dimensional random functions in the CLT regime**). *For random variable $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n}\cdot(\|\mathbf{x}\|^2 - 1)$, with $\sqrt{n}\mathbf{x} \in \mathbb{R}^n$ having i.i.d. standard Gaussian entries, in the CLT regime (as in the second item of Example 3.2), it follows from the CLT that $f_{\text{CLT}}(\mathbf{x}) \sim \mathcal{N}(0,1)$ in law as $n \to \infty$. Thus, it follows from Theorem 3.10 that*

$$\mathbb{E}[\phi(\sqrt{n}\cdot(\|\mathbf{x}\|^2 - 1))] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi(x)] + o(1) = a_{\phi;0} + o(1), \tag{3.27}$$

with high probability, where $o(1)$ denotes quantity that goes to zero as $n \to \infty$. Similarly, for random variable $f_{\text{CLT}}(\mathbf{x}) = \sqrt{n}\cdot\mathbf{x}^\mathsf{T}\mathbf{y}$ with $\|\mathbf{y}\| = 1$, it follows that

$$\mathbb{E}[\phi(\sqrt{n}\cdot\mathbf{x}^\mathsf{T}\mathbf{y})] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\phi(x)] = a_{\phi;0}, \tag{3.28}$$

where we do not have the error term $o(1)$ since $\sqrt{n}\cdot\mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0,1)$ for any *n.*

Proposition 3.12 presents the high-dimensional linearization in the CLT regime via the Hermite polynomial expansion. This approach should be compared and contrasted with that of Proposition 3.6, which presents the high-dimensional linearization in the LLN regime via the Taylor expansion method. The distinction between these two methodologies is elaborated upon in the following remark.

The idea of orthogonal polynomials in Definition 3.8 and Theorem 3.10 applies to other nonlinear forms beyond the simple expectation $\mathbb{E}[\phi(x)]$. In particular, it applies to nonlinear forms that involve large-dimensional random vectors and matrices. See Chapter 3.4 below for an in-depth discussion on its use in assessing nonlinear random vectors and **??** for an exposition with applications to ML.

Remark 3.13 (**Different scalings, Taylor expansion versus orthogonal polynomial**). We can compare and contrast the two linearization approaches of Taylor expansion (in Theorem 3.5) and orthogonal Hermite polynomial expansion (in Theorem 3.10), to assess the nonlinear objects in Example 3.2 in the LLN and CLT regimes, respectively. Recall from Example 3.2 that for a random vector $\mathbf{x} \in \mathbb{R}^n$ such that $\sqrt{n}\mathbf{x}$ has i.i.d. standard Gaussian entries, a deterministic $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\|_2 = 1$, we have $\mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0, n^{-1})$ so that

$$\begin{aligned} f_{\text{LLN}}(\mathbf{x}) &\equiv \mathbf{x}^\mathsf{T}\mathbf{y} = 0 + O(n^{-1/2}), \\ f_{\text{CLT}}(\mathbf{x}) &\equiv \sqrt{n}\cdot\mathbf{x}^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0,1). \end{aligned}$$

We are interested in the behavior of $\phi(f_{\text{LLN}}(\mathbf{x}))$ and $\phi(f_{\text{CLT}}(\mathbf{x}))$, and in particular, how they depend on the nonlinear $\phi\colon \mathbb{R} \to \mathbb{R}$. We have the following.

1. **In the LLN regime**, by Taylor expansion (of nonlinear LLN random variables) in Proposition 3.6, any pair of smooth function ϕ, ψ with $\phi(0) = \psi(0)$ satisfies

$$\phi(f_{\mathrm{LLN}}(\mathbf{x})) = \psi(f_{\mathrm{LLN}}(\mathbf{x})) + O(n^{-1/2}), \tag{3.29}$$

   with high probability for n large. Thus, the two random variables, $\phi(f_{\mathrm{LLN}}(\mathbf{x}))$ and $\psi(f_{\mathrm{LLN}}(\mathbf{x}))$, are close as long as the two nonlinear functions ϕ and ψ coincide at 0.

2. **In the CLT regime**, by Hermite polynomial expansion in Proposition 3.12 for ϕ, ψ having the same *zeroth-order* Hermite coefficient $a_{\phi;0} = \mathbb{E}[\phi(x)] = a_{\psi;0} = \mathbb{E}[\psi(x)]$ with $x \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[\phi(f_{\mathrm{CLT}}(\mathbf{x}))] = \mathbb{E}[\psi(f_{\mathrm{CLT}}(\mathbf{x}))]. \tag{3.30}$$

   This is by no means surprising, as it is a consequence of the definition $a_{\phi;0} = \mathbb{E}[\phi(x)] = a_{\psi;0} = \mathbb{E}[\psi(x)]$.

In order to understand Remark 3.13 better, we provide in the following a concrete example of the two linearization approaches.

Example 3.14 (**Two different linearizations of** tanh **in two different scaling regimes**)**.** *As a concrete example of Remark 3.13, consider the hyperbolic tangent function $\phi(t) = \tanh(t)$. It follows from the discussions in Remark 3.13 that this nonlinear function is "close" to different quadratic functions in different regimes of interest. More precisely, for a random vector $\mathbf{x} \in \mathbb{R}^n$ such that $\sqrt{n}\mathbf{x}$ has i.i.d. standard Gaussian entries, a deterministic $\mathbf{y} \in \mathbb{R}^n$ of unit norm $\|\mathbf{y}\|_2 = 1$, we have the following.*

1. ***In the LLN regime**, we have for $f_{\mathrm{LLN}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{y}$ that*

$$\tanh(f_{\mathrm{LLN}}(\mathbf{x})) \simeq \psi_{\mathrm{LLN}}(f_{\mathrm{LLN}}(\mathbf{x})),$$

   with $\psi_{\mathrm{LLN}}(x) = x^2/4$. This is as a consequence of $\tanh(x = 0) = \psi_{\mathrm{LLN}}(x = 0) = 0$. In particular, we also have $\mathbb{E}[\tanh(f_{\mathrm{LLN}}(\mathbf{x}))] \simeq \mathbb{E}[\psi(f_{\mathrm{LLN}}(\mathbf{x}))]$ as a result.

2. ***In the CLT regime**, we have for $f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\mathsf{T}\mathbf{y}$ that*

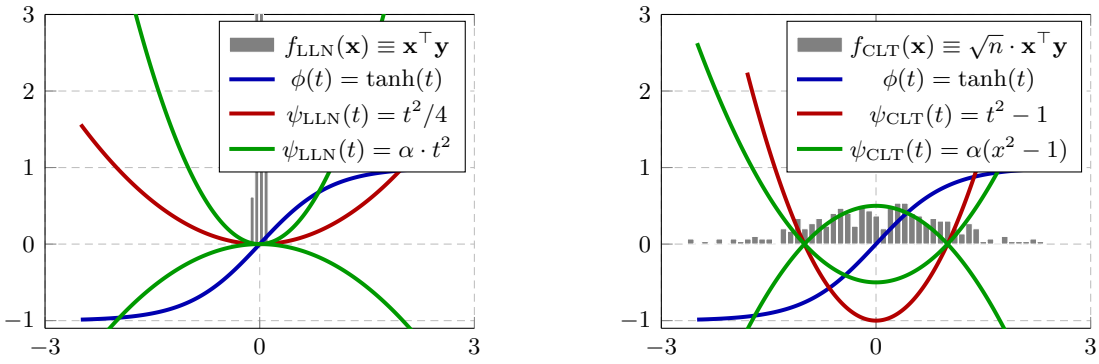
$$\mathbb{E}[\tanh(f_{\mathrm{CLT}}(\mathbf{x}))] = \mathbb{E}[\psi_{\mathrm{CLT}}(f_{\mathrm{CLT}}(\mathbf{x}))]$$

   in expectation, with now $\psi_{\mathrm{LLN}}(x) = x^2 - 1$, i.e., with a different function. This is a consequence of the fact that their zeroth-order Hermite coefficient $a_0 = 0$.

Figure 3.4 visually compares the behavior of $\tanh(f_{\mathrm{LLN}}(\mathbf{x}))$ and $\tanh(f_{\mathrm{CLT}}(\mathbf{x}))$, in the LLN and CLT regime.

3.4 Linearization of $f(\phi(\mathbf{x}))$ with Linear Equivalent

In this section, we discuss how the linearization techniques (of Taylor and orthogonal polynomial expansions) for *scalar variables* extend to *multivariate vector variables* $\phi(\mathbf{x})$ for some $\phi\colon \mathbb{R}^n \to \mathbb{R}^n$ that applies entry-wise on the random vector $\mathbf{x} \in \mathbb{R}^n$, when their *scalar observations* of the form $f(\phi(\mathbf{x}))$ are considered, as in the bottom half of Table 3.1. Recall that Example 3.2 and Chapters 3.2 and 3.3 focus on Taylor expansion and orthogonal polynomial expansion for *scalar nonlinear* random variables of the form $\phi(f(\mathbf{x}))$ for $f\colon \mathbb{R}^n \to \mathbb{R}$ (such as inner products and norms of vectors in Example 3.2) and $\phi\colon \mathbb{R} \to \mathbb{R}$, in the two different LLN and CLT scaling regimes. Here, we show that these two technical approaches extend beyond the case of scalar nonlinear random variables like $\phi(f(\mathbf{x}))$ to *nonlinear random vectors* $\phi(\mathbf{x})$ with entry-wise ϕ,



(a) LLN regime (b) CLT regime

Figure 3.4: Different behavior of nonlinear $\phi(f_{\text{LLN}}(\mathbf{x}))$ and $\phi(f_{\text{CLT}}(\mathbf{x}))$ for $\phi(t) = \tanh(t)$ (in **blue**) in the LLN and CLT regime, with $n = 500$. We have $\phi(f_{\text{LLN}}(\mathbf{x})) \simeq \psi_{\text{LLN}}(f_{\text{LLN}}(\mathbf{x}))$ in the LLN regime (as a consequence of $\phi(0) = \psi_{\text{LLN}}(0) = 0$) and $\mathbb{E}[\phi(f_{\text{CLT}}(\mathbf{x}))] = \mathbb{E}[\psi_{\text{CLT}}(f_{\text{CLT}}(\mathbf{x}))]$ in the CLT regime (as a consequence of $a_{\phi;0} = a_{\psi_{\text{CLT}};0} = 0$), with *different* quadratic functions $\psi_{\text{LLN}}(t) = t^2/4$ and $\psi_{\text{CLT}}(t) = t^2 - 1 = \sqrt{2}\text{He}_2(t)$ in **red**. Note that the these linearizations (in the two different regimes respectively) are *not* unique and all functions in dashed **green** are also valid linearizations.

and in particular, their scalar observations $f(\phi(\mathbf{x}))$ via some $f\colon \mathbb{R}^n \to \mathbb{R}$. This can be done by studying the associated Linear Equivalent, defined as follows.

Linear Equivalent

Definition 3.15 (Linear Equivalent). *For a random vector $\mathbf{x} \in \mathbb{R}^n$, its nonlinear transformation $\phi(\mathbf{x}) \in \mathbb{R}^n$ is obtained by applying $\phi\colon \mathbb{R}^n \to \mathbb{R}^n$ entry-wise on $\mathbf{x}$. Consider $f(\phi(\mathbf{x}))$ a* scalar *observation of $\phi(\mathbf{x}) \in \mathbb{R}^n$ via observation function $f\colon \mathbb{R}^n \to \mathbb{R}$, we say that the random vector $\tilde{\mathbf{x}}_\phi$ (defined on an extended probability space if necessary) is an* (ε, δ)-Linear Equivalent *of the nonlinear $\phi(\mathbf{x})$ if, with probability at least $1 - \delta(n)$ that*

$$|f(\phi(\mathbf{x})) - f(\tilde{\mathbf{x}}_\phi)| \le \varepsilon(n), \tag{3.31}$$

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

$$f(\phi(\mathbf{x})) - f(\tilde{\mathbf{x}}_\phi) \to 0, \tag{3.32}$$

in probability or almost surely, and we denote

$$\phi(\mathbf{x}) \overset{f}{\leftrightarrow} \tilde{\mathbf{x}}_\phi. \tag{3.33}$$

The Linear Equivalent in Definition 3.15 is a special case of the High-dimensional Equivalent in Definition 1.1 for vectors.

As expected, the nonlinear object of interest, as well as the corresponding Linear Equivalent in Definition 3.15, depends on whether the nonlinear $\phi\colon \mathbb{R} \to \mathbb{R}$ is applied to (the entries of) the random vector $\mathbf{x}$ in the LLN or the CLT regime (see Definition 3.1), as illustrated in Example 3.2 for $\phi(f_{\text{LLN}}(\mathbf{x}))$ versus $\phi(f_{\text{CLT}}(\mathbf{x}))$. We should resort to the Taylor expansion in Chapter 3.2 for the LLN regime and the orthogonal polynomial expansion approach in Chapter 3.3 for CLT regime, respectively.

In Algorithm 1 and 2, we present algorithms based on Taylor and Hermite polynomial expansions discussed in Chapter 3.2 and Chapter 3.3, to construct Linear Equivalent for $f(\phi(\mathbf{x}))$, in the LLN and CLT regime, respectively.

Algorithm 1: Linear Equivalent for $f(\phi_{\mathrm{LLN}}(\mathbf{x}))$ in the LLN regime

Input: Nonlinear random vector $\phi_{\mathrm{LLN}}(\mathbf{x}) \in \mathbb{R}^n$ in the LLN regime so that the entries of $\mathbf{x}$ satisfy $x_i \approx \tau$ for $i \in \{1, \ldots, n\}$ and its scalar observation $f(\phi_{\mathrm{LLN}}(\mathbf{x}))$ of interest.

Output: Linear Equivalent $\tilde{\mathbf{x}}_{\phi_{\mathrm{LLN}}} \overset{f}{\leftrightarrow} \phi(\mathbf{x})$ when $f(\phi_{\mathrm{LLN}}(\mathbf{x}))$ is considered.

In the LLN regime, call Theorem 3.5 to linearize the i^{th} entry of $\phi_{\mathrm{LLN}}(\mathbf{x})$ as

$$\phi_{\mathrm{LLN}}(x_i) = \phi_{\mathrm{LLN}}(\tau) + \phi'_{\mathrm{LLN}}(\tau)(x_i - \tau) + \frac{1}{2}\phi'_{\mathrm{LLN}}(\tau)(x_i - \tau)^2 + \ldots + \varepsilon,$$

to some desired linearization error ε;

return $\tilde{\mathbf{x}}_{\phi_{\mathrm{LLN}}} = \phi_{\mathrm{LLN}}(\tau) \cdot \mathbf{1}_n + \phi'_{\mathrm{LLN}}(\tau)(\mathbf{x} - \tau \cdot \mathbf{1}_n) + \ldots$ *such that with high probability* $\|\phi_{\mathrm{LLN}}(\mathbf{x}) - \tilde{\mathbf{x}}_{\phi}\|_\infty = \varepsilon$.

Note that

1. in the LLN regime in Algorithm 1, the linearization and corresponding Linear Equivalent of $\phi_{\mathrm{LLN}}(\mathbf{x})$ only depend on the entry-wise non-linearity ϕ_{LLN}, in particular its local behavior around the point of LLN-concentration τ;
2. in the CLT regime in Algorithm 2, the linearization and corresponding Linear Equivalent of $\phi_{\mathrm{CLT}}(\mathbf{x})$ depend on
   (a) the distribution of the random vector $\mathbf{x}$ (which determines the family of orthogonal polynomials, see, e.g., Remark 3.11 for a discussion); and
   (b) the number (and in fact form of the one or more) of scalars observations $f_i(\cdot)$ of the nonlinear $\phi_{\mathrm{CLT}}(\mathbf{x})$.

In the following, we extend the scalar nonlinear objects in Example 3.2 to scalar observations of nonlinear random vectors, in both the LLN and the CLT regime.

Also, note that Example 3.16 in NOT in perfect parallel to Example 3.2, since here we only consider the inner product as objects, so let us discuss.

Example 3.16 (**Scalar observations of 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 (the scaling by* $\sqrt{n}$ *is made so that the rows* $\mathbf{x}_i^{\mathsf{T}} \in \mathbb{R}^{1 \times n}$ *of* $\mathbf{X}$ *satisfy* $\mathbb{E}[\|\mathbf{x}_i\|]_2^2 = 1$ *as in Example 3.2 and Remark 2.4), and* $\mathbf{y} \in \mathbb{R}^n, \mathbf{a} \in \mathbb{R}^p$ *be* deterministic *vectors of unit norm* $\|\mathbf{y}\| = 1, \|\mathbf{a}\| = 1$*; and consider the following* scalar observations *of* nonlinear *random vectors with observation function* $f \colon \mathbb{R}^p \to \mathbb{R}$ *and entry-wise nonlinear function* $\phi \colon \mathbb{R}^p \to \mathbb{R}^p$ *acting in two different regimes:*

1. ***LLN regime****: here, we consider* $f(\phi_{\mathrm{LLN}}(\mathbf{X}\mathbf{y})) = \mathbf{a}^{\mathsf{T}}\phi_{\mathrm{LLN}}(\mathbf{X}\mathbf{y})/\sqrt{p}$*; and*
2. ***CLT regime****:* $f(\phi_{\mathrm{CLT}}(\sqrt{n} \cdot \mathbf{X}\mathbf{y})) = \mathbf{a}^{\mathsf{T}}\phi_{\mathrm{CLT}}(\sqrt{n} \cdot \mathbf{X}\mathbf{y})/\sqrt{p}$*,*

where we consider the scalar observation $f(\cdot) = \mathbf{a}^{\mathsf{T}}(\cdot)/\sqrt{p}, \|\mathbf{a}\|_2 = 1$ *as an illustrating example, among those thoroughly discussed in Chapter 1.*

Remark 3.17 (**Example 3.16 versus Example 3.2**). Comparing Example 3.16 for vectors to Example 3.2 for scalars, we remark that:

- Here in Example 3.16, the i^{th} entry of the (entry-wise) nonlinear random vector $\phi_{\mathrm{LLN}}(\mathbf{X}\mathbf{y})$ and $\phi_{\mathrm{CLT}}(\mathbf{X}\mathbf{y})$ is nothing but a scalar non-linearity $\phi \colon \mathbb{R} \to \mathbb{R}$ acting respectively on scalars $\mathbf{x}^{\mathsf{T}}\mathbf{y}$ and $\sqrt{n} \cdot \mathbf{x}^{\mathsf{T}}\mathbf{y}$ in the LLN and CLT regime as in Example 3.2.

Algorithm 2: Linear Equivalent of $f(\phi_{\text{CLT}}(\mathbf{x}))$ in the CLT regime

Input: Nonlinear random vector $\phi_{\text{CLT}}(\mathbf{x}) \in \mathbb{R}^n$ in the CLT regime with, e.g., standard Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and its K scalar observations $f_1(\phi_{\text{CLT}}(\mathbf{x})), f_2(\phi_{\text{CLT}}(\mathbf{x})), \ldots, f_K(\phi_{\text{CLT}}(\mathbf{x}))$ of interest.

Output: Equivalent $\tilde{\mathbf{x}}_{\phi_{\text{CLT}}} \overset{f_1,\ldots,f_K}{\longleftrightarrow} \phi(\mathbf{x})$ when the *joint behavior* of $f_1(\phi_{\text{CLT}}(\mathbf{x})), \ldots, f_K(\phi_{\text{CLT}}(\mathbf{x}))$ is considered.

Initialize $\tilde{\mathbf{x}}_{\phi_{\text{CLT}}} \leftarrow \mathbf{0}_n$;

for $i = 1$ **to** K **do**

  if $f_i(\phi_{\text{CLT}}(\mathbf{x})) \neq \|\phi_{\text{CLT}}(\mathbf{x})\|^2/n$ *is* not *the (squared normalized) norm of* $\phi_{\text{CLT}}(\mathbf{x})$ **then**

    introduce the i^{th} Hermite polynomial $P_i(\mathbf{x})$ of $\mathbf{x}$ as defined in Equation (3.21) of Theorem 3.10;

    determine the corresponding coefficient $\alpha_{\phi;i}$ via Hermite polynomial expansion of ϕ_{CLT} as in Equation (3.24), so that $f_i\left(\tilde{\mathbf{x}}_{\phi_{\text{CLT}}} + \alpha_{\phi;i} P_i(\mathbf{x})\right) \simeq f_i(\phi_{\text{CLT}}(\mathbf{x}))$;

    set $\tilde{\mathbf{x}}_{\phi_{\text{CLT}}} \leftarrow \tilde{\mathbf{x}}_{\phi_{\text{CLT}}} + \alpha_{\phi;i} P_i(\mathbf{x})$;

  else

    introduce a *fresh* random vector $\mathbf{z} \in \mathbb{R}^p$ having i.i.d. standard Gaussian entries and *independent* of $\mathbf{x}$;

    determine the corresponding coefficient β so that $f_i(\tilde{\mathbf{x}}_{\phi_{\text{CLT}}} + \beta\mathbf{z}) \simeq f_i(\phi_{\text{CLT}}(\mathbf{x}))$ for $f_i(\phi_{\text{CLT}}(\mathbf{x})) \neq \|\phi_{\text{CLT}}(\mathbf{x})\|^2/n$, by setting $\beta = \sqrt{\nu_\phi - \mathbb{E}[\tilde{\mathbf{x}}_{\text{CLT}}^{\mathsf{T}} \tilde{\mathbf{x}}_{\text{CLT}}]/n}$ with ν_ϕ defined in Equation (3.25);

    set $\tilde{\mathbf{x}}_{\phi_{\text{CLT}}} \leftarrow \tilde{\mathbf{x}}_{\phi_{\text{CLT}}} + \beta\mathbf{z}$;

  end

end

- While in Example 3.16 we focus on inner products (between the rows of $\mathbf{X}$ and $\mathbf{y}$), norms (as in Example 3.2) can be studied similarly.

- Different from Example 3.2 where the randomness comes from the vector $\mathbf{x} \in \mathbb{R}^n$, here in Example 3.16 the randomness comes from the *matrix* $\mathbf{X} \in \mathbb{R}^{p \times n}$ that involves two dimensions n and p. Intuitively, the dimension n plays the same role as in Example 3.2, and leads to LLN- or CLT-type concentration of the entries of $\mathbf{Xy}$ or $\sqrt{n} \cdot \mathbf{Xy}$, on which ϕ_{LLN} or ϕ_{CLT} is applied; on the other hand, the dimension p should also be large, so that the scalar observation $f(\cdot) = \mathbf{a}^{\mathsf{T}}(\cdot)/\sqrt{p}$ concentrates, as discussed in Chapter 1. (In this sense, the scalar observation $f(\cdot)$ is chosen so that it establish LLN-type concentration.) So, here in Example 3.16 we are working in the RMT *proportional regime* as $n, p \to \infty$ together.

We describe next how the Taylor expansion approach in Theorem 3.5 and the orthogonal Hermite polynomial expansion approach in Theorem 3.10 discussed in previous sections apply to linearize (the observations of) the nonlinear random vector $\phi(\mathbf{Xy})$ and get the corresponding Linear Equivalent in Definition 3.15.

Taylor expansion for Linear Equivalent in the LLN regime. We first evaluate the scalar observation $f(\cdot) = \mathbf{a}^{\mathsf{T}}(\cdot)/\sqrt{p}$ of the nonlinear random vector $\phi_{\text{LLN}}(\mathbf{Xy})$ in the LLN regime, as in the first item of Example 3.16. Its corresponding Linear Equivalent can be obtained using Taylor expansion in Theorem 3.5 and is given in the following result.

Proposition 3.18 (**Linear Equivalent 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*

$\mathbf{y} \in \mathbb{R}^n, \mathbf{a} \in \mathbb{R}^p$ *be* deterministic *vectors of unit norm* $\|\mathbf{y}\| = 1, \|\mathbf{a}\| = 1$*, the following Linear Equivalent (Definition 3.15) holds in the LLN regime (as in the first item of Example 3.16):*

$$\phi_{\text{LLN}}(\mathbf{X}\mathbf{y}) \overset{f}{\leftrightarrow} \underbrace{\phi_{\text{LLN}}(0) \cdot \mathbf{1}_p}_{O_{\|\cdot\|_\infty}(1)} + \underbrace{\phi'_{\text{LLN}}(0) \cdot \mathbf{X}\mathbf{y}}_{O_{\|\cdot\|_\infty}(n^{-1/2})}, \tag{3.34}$$

for scalar observation $f(\cdot) = \mathbf{a}^\mathsf{T}(\cdot)/\sqrt{p}$*, up to some error of order* $o(1/\sqrt{np})$*.*

Proof of Proposition 3.18. To prove Proposition 3.18, note that for $\mathbf{y}$ of unit norm and $\mathbf{X}$ having i.i.d. Gaussian entries of mean zero and variance $1/n$, we have that the entries of $\mathbf{X}\mathbf{y}$ are i.i.d. Gaussian of mean zero and variance $1/n$, so that

$$\|\mathbf{X}\mathbf{y}\|_\infty = O(n^{-1/2}), \tag{3.35}$$

with high probability for n large. As such, the nonlinear ϕ_{LLN} applied on the entries of $\mathbf{X}\mathbf{y}$ in the LLN regime, with point of LLN-concentration $\tau = 0$. We then proceed as in Algorithm 1, to Taylor expand, for n large, the i^{th} entry of the nonlinear random vector $\phi_{\text{LLN}}(\mathbf{X}\mathbf{y})$ as

$$\phi_{\text{LLN}}(\mathbf{x}_i^\mathsf{T}\mathbf{y}) = \phi_{\text{LLN}}(0) + \phi'_{\text{LLN}}(0)(\mathbf{x}_i^\mathsf{T}\mathbf{y}) + O(n^{-1}), \tag{3.36}$$

where we denote $\mathbf{x}_i^\mathsf{T} \in \mathbb{R}^{1\times n}$ the i^{th} row of $\mathbf{X} \in \mathbb{R}^{p\times n}$. This leads to the infinity norm approximation of $\phi_{\text{LLN}}(\mathbf{X}\mathbf{y})$ as

$$\phi_{\text{LLN}}(\mathbf{X}\mathbf{y}) = \underbrace{\phi_{\text{LLN}}(0) \cdot \mathbf{1}_p}_{O_{\|\cdot\|_\infty}(1)} + \underbrace{\phi'_{\text{LLN}}(0) \cdot \mathbf{X}\mathbf{y}}_{O_{\|\cdot\|_\infty}(n^{-1/2})} + O_{\|\cdot\|_\infty}(n^{-1}), \tag{3.37}$$

for $O_{\|\cdot\|_\infty}(n^{-1})$ a vector having infinity norm of order $O(n^{-1})$ with high probability. As such, we have, for the scalar observation $f(\cdot) = \mathbf{a}^\mathsf{T}(\cdot)/\sqrt{p}$ of $\phi(\mathbf{X}\mathbf{y})$ that

$$\mathbf{a}^\mathsf{T}\phi_{\text{LLN}}(\mathbf{X}\mathbf{y})/\sqrt{n} = \phi_{\text{LLN}}(0)\underbrace{\boldsymbol{\alpha}^\mathsf{T}\mathbf{1}_p/\sqrt{p}}_{O(1)} + \phi'_{\text{LLN}}(0)\underbrace{\boldsymbol{\alpha}^\mathsf{T}\mathbf{X}\mathbf{y}/\sqrt{p}}_{O(1/\sqrt{np})} + o(1/\sqrt{np}), \tag{3.38}$$

where we used the fact that $\boldsymbol{\alpha}^\mathsf{T}\mathbf{X}\mathbf{y} = \sum_{i=1}^p \sum_{j=1}^n \alpha_i y_j X_{ij} \sim \mathcal{N}(0, n^{-1})$ as the weighted sum of np independent Gaussian random variables. This concludes the proof of Proposition 3.18. □

Hermite polynomial expansion for Linear Equivalent in the CLT regime. Now, we evaluate the (same) scalar observation $f(\cdot) = \mathbf{a}^\mathsf{T}(\cdot)/\sqrt{p}$ as above, but of the nonlinear random vector $\phi_{\text{CLT}}(\sqrt{n} \cdot \mathbf{X}\mathbf{y})$ in the CLT regime, as in the second item of Example 3.16. Its corresponding Linear Equivalent can be obtained using Hermite polynomial expansion in Theorem 3.10 and is given in the following result.

Proposition 3.19 (Linear Equivalent 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* $\mathbf{y} \in \mathbb{R}^n, \mathbf{a} \in \mathbb{R}^p$ *be* deterministic *vectors of unit norm* $\|\mathbf{y}\| = 1, \|\boldsymbol{\alpha}\| = 1$*, the following Linear Equivalent (Definition 3.15) holds in the CLT regime (as in the second item of Example 3.16):*

$$\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y}) \overset{f}{\leftrightarrow} a_{0,\phi} \cdot \mathbf{1}_p, \tag{3.39}$$

for scalar observation $f(\cdot) = \mathbf{a}^\mathsf{T}(\cdot)/\sqrt{p}$*, up to some error of order* $O(p^{-1/2}$*.*

Proof of Proposition 3.19. To prove Proposition 3.19, note that for $\mathbf{X} \in \mathbb{R}^{p\times n}$ with i.i.d. standard Gaussian entries with zero mean and variance n^{-1}, and $\mathbf{y} \in \mathbb{R}^n$ of unit norm, the random vector $\sqrt{n}\mathbf{X}\mathbf{y} \in \mathbb{R}^p$ has standard i.i.d. Gaussian entries of zero mean and unit variance, so that

the nonlinear ϕ_{CLT} applied on the entries of $\sqrt{n}\mathbf{X}\mathbf{y}$ in the CLT regime. We then proceed as in Algorithm 2. It follows from Theorem 3.10 that, for the i^{th} entry of the nonlinear random vector $\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})$, we have the following formal expansion

$$\phi_{\text{CLT}}(\sqrt{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{y}) \sim \sum_{\ell=0}^{\infty} a_{\phi;\ell}\text{He}_\ell(\sqrt{n}\mathbf{x}_i^{\mathsf{T}}\mathbf{y}), \tag{3.40}$$

where we denote $\mathbf{x}_i^{\mathsf{T}} \in \mathbb{R}^{1\times n}$ the i^{th} row of $\mathbf{X}$, and $a_{\phi;\ell}$ the ℓ^{th} Hermite coefficient of ϕ_{CLT}.

At this point, note from Equation (3.40) that the approximation of (the i^{th} entry of) the nonlinear random vector $f(\sqrt{n}\mathbf{X}\mathbf{y})$ in the CLT regime with Hermite polynomial is only "accurate" as the degree $L \to \infty$. As such, the direct accurate approximation of ϕ using the orthogonal polynomial framework comes at the cost of computing a large (or even an infinite) number of coefficients $a_{\phi;\ell}$. While it is possible to simplify such approximation by making additional regularity assumption on ϕ_{CLT} so that, e.g., the coefficients $a_{\phi;\ell}$ decay sufficiently fast as ℓ grows large and that the higher-orders terms can be ignored in the approximation, not much more can be said in the general case, for the nonlinear random vector $\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})$.

On the other hand, recall from Proposition 3.12 that, (very) simple Hermite polynomial expansion exists for the *expectation* $\mathbb{E}[\phi_{\text{CLT}}(\sqrt{n}\mathbf{x}^{\mathsf{T}}\mathbf{y})]$ of the nonlinear random variable $\phi_{\text{CLT}}(\sqrt{n}\mathbf{x}^{\mathsf{T}}\mathbf{y})$, which depends *only* on the zeroth-order Hermite coefficient of ϕ_{CLT}. This, together with the fact that scalar observations (at least those discussed in Chapter 1, including the linear map $f(\cdot) = \mathbf{a}^{\mathsf{T}}(\cdot)/\sqrt{p}$) of large-dimensional random vectors concentrate around their expectations, allows one to prove Proposition 3.19.

Precisely, recall that $\sqrt{n}\mathbf{X}\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, it follows the LLN and CLT that

$$\begin{aligned} f(\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})) &= \mathbf{a}^{\mathsf{T}}\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})/\sqrt{p} = \underbrace{\mathbf{a}^{\mathsf{T}}\mathbb{E}[\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})]/\sqrt{p}}_{O(1)} + O(p^{-1/2}) \\ &= a_{\phi;0} \cdot \underbrace{\mathbf{a}^{\mathsf{T}}\mathbf{1}_p/\sqrt{p}}_{O(1)} + O(p^{-1/2}), \end{aligned}$$

with high probability for p large.[8] This concludes the proof of Proposition 3.19. □

Remark 3.20 (**Proposition 3.18 versus Proposition 3.19.**)**.** Comparing the Linear Equivalents in Proposition 3.18 in the LLN regime to Proposition 3.19 in the CLT regime, we observe the following.

- The two Linear Equivalents are *similar*, in that they are (in their first order) both proportional to the vector of all ones.
- The two results are *different*, in that:
  1. Proposition 3.18 performs a local Taylor expansion of ϕ_{LLN} in the LLN regime around 0, the point of LLN-concentration, so that the obtained Linear Equivalent depends on ϕ_{LLN} only via its *local* behavior, while that in Proposition 3.19 in the CLT regime is obtained via the Hermite polynomial expansion, and depends on the *global* behavior of ϕ_{CLT} (e.g., via its zeroth order Hermite coefficient $a_{\phi;0}$); and
  2. the form of the Linear Equivalent in the LLN regime in Proposition 3.18 is *independent* of the observation $f(\cdot)$ (note that the derivation of Proposition 3.18 holds for *any* $f(\cdot)$), while that in the CLT regime in Proposition 3.19 and Algorithm 2 relies on the computation of the expectation of $\mathbb{E}[f(\cdot)]$ and thus *depends* on the form of f. See Example 3.21 below for an example.

[8] Note that the LLN-type concentration of $f(\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})) = \mathbb{E}[f(\phi_{\text{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y}))] + O(p^{-1/2})$ is *not* a consequence of the Hermite polynomial expansion, and needs be proven separately using, e.g., LLN and CLT.

Example 3.21 (**Linear Equivalent in the CLT regime: random observation function**). *Under the same notations and settings as in Proposition 3.19 but for* random *observation function*

$$f(\cdot) = \mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}(\cdot)/\sqrt{n}, \tag{3.41}$$

that is assumed to establish LLN-type concentrate around its expectation up to some error $\varepsilon(n,p)$ *for* n, p *large*[9]*, the following Linear Equivalent (Definition 3.15) holds in the CLT regime (as in the second item of Example 3.16):*

$$\phi_{\mathrm{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y}) \overset{f}{\leftrightarrow} a_{\phi;1} \cdot \sqrt{n}\mathbf{X}\mathbf{y}. \tag{3.42}$$

Proof of Example 3.21. To prove Example 3.21, note that $\sqrt{n}\mathbf{X}\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ in the CLT regime, so that by the (assumption of) LLN-type concentration , it remains to compute the following expectation

$$\mathbb{E}[f(\phi_{\mathrm{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y}))] = \frac{1}{\sqrt{n}}\mathbb{E}[\mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}\phi_{\mathrm{CLT}}(\sqrt{n}\mathbf{X}\mathbf{y})] = \frac{1}{n}\sum_{i=1}^{p}\mathbb{E}[(\sqrt{n}\mathbf{x}_i^\mathsf{T}\mathbf{y})\phi_{\mathrm{CLT}}(\sqrt{n}\mathbf{x}_i^\mathsf{T}\mathbf{y})] = \frac{p}{n}a_{\phi;1}, \tag{3.43}$$

with high probability up to some error $\varepsilon(n,p)$, for $\sqrt{n}\mathbf{x}_i^\mathsf{T}\mathbf{y} \sim \mathcal{N}(0,1)$. Note that this Linear Equivalent is different from that in Proposition 3.19). Similarly, we have

$$\frac{1}{\sqrt{n}}\mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}\left(a_{1,f}\cdot\sqrt{n}\mathbf{X}\mathbf{y}\right) = a_{1,f}\cdot\mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{X}^\mathsf{T}\mathbf{X}]\mathbf{y} + O\left(\frac{\sqrt{p}}{n}\right) = \frac{p}{n}a_{\phi;1} + O\left(\frac{\sqrt{p}}{n}\right), \tag{3.44}$$

where the error term $O(\sqrt{p}/n)$ arises due to the following concentration of $\mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{y}$:

$$\mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{y} = \sum_{i=1}^{p}(\mathbf{y}^\mathsf{T}\mathbf{x}_i)^2 = \sum_{i=1}^{p}\mathbb{E}[(\mathbf{y}^\mathsf{T}\mathbf{x}_i)^2] + O\left(\sqrt{p}(\mathbf{y}^\mathsf{T}\mathbf{x}_i)^2\right) = \mathbf{y}^\mathsf{T}\mathbb{E}[\mathbf{X}^\mathsf{T}\mathbf{X}]\mathbf{y} + O\left(\frac{\sqrt{p}}{n}\right), \tag{3.45}$$

per the LLN and CLT, where $\mathbf{x}_i^\mathsf{T} \in \mathbb{R}^{1\times n}$ is the i^{th} row of $\mathbf{X} \in \mathbb{R}^{p\times n}$. □

The fact that in the CLT regime, linearization and the corresponding Linear Equivalent depend on the observation function, as we shall see in **??**, plays a crucial role in linearizing nonlinear random matrices.

[9]Note that this concentration result looks like, but is formally different from that of quadratic or nonlinear quadratic forms in Theorem 1.22 and Theorem 1.24, and needs be proven separately.

Part II

Four ways to characterize sample covariance matrices

In this Part, we move on to consider the behavior of random matrices, starting with the fundamental object of the sample covariance matrix (SCM).[10] Let's say we are given n independent centered data samples, $\mathbf{x}_i \in \mathbb{R}^p$, with $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}_p$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{C}$. From this, one can construct a data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, the SCM of which is defined as follows.

Sample Covariance Matrix (SCM)

Definition 4.22 (**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*

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \frac{1}{n} \mathbf{X}\mathbf{X}^\mathsf{T}. \tag{4.46}$$

Depending on the dimensionality of n and p, we introduce two different scaling regimes, the classical regime and the proportional regime,[11] defined in the context of SCMs as follows.

Classical versus proportional regimes

Definition 4.23 (**Classical versus proportional regimes**). *For a SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p \times p}$ computed from n samples of dimension p, as in Definition 4.22, we 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 ($n \gg p$ for large but finite n) characterizations.*
2. ***Proportional regime**: with $n \sim p$; this includes both asymptotic ($n, p \to \infty$ with $p/n \to c \in (0, \infty)$ and non-asymptotic ($n \sim p \gg 1$ both large but finite) characterizations.*

We will present different ways to characterize the (spectral) behavior of a SCM:

1. by considering the **classical** ($n \gg p$) as well as the **proportional** ($n \sim p$) regimes; and
2. by providing **asymptotic** (as $n \to \infty$ and/or $p \to \infty$) as well as **non-asymptotic** (for n, p large but finite) guarantees.

[10] Among other things, in the case that $\mathbf{x}_i$ follows a multivariate Gaussian distribution with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, the SCM is the maximum likelihood estimator [1] of the population covariance $\mathbf{C}$.

[11] The proportional regime is sometimes known as the *thermodynamic limit* in the statistical physics literature [28, 39].

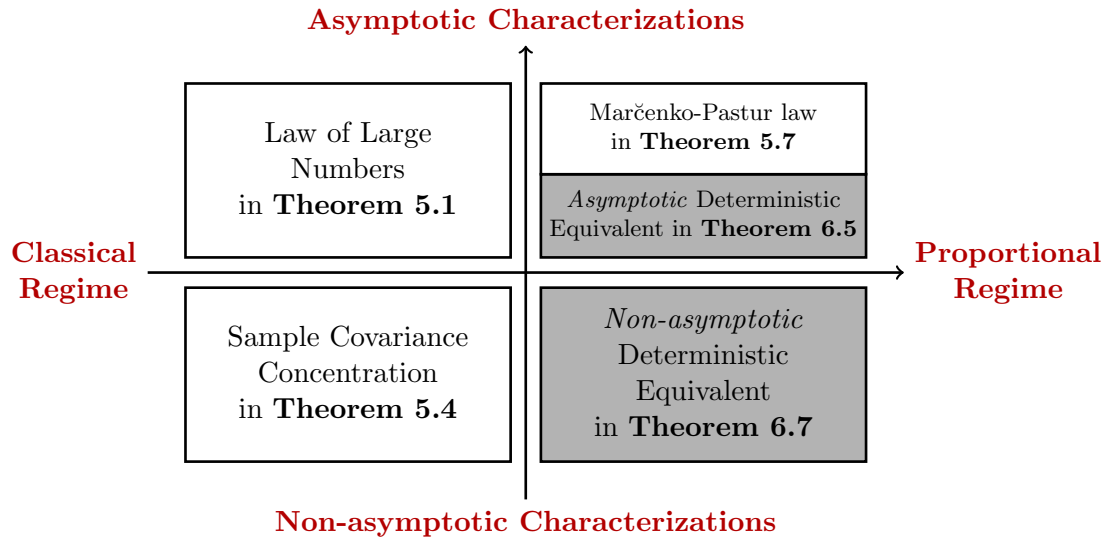


Figure 4.5: Taxonomy of four different ways to characterize the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, depending on whether one works in the classical regime versus proportional regime, and whether one is interested in asymptotic results or non-asymptotic results. More traditional "old school" results are in white (see Chapter 5), while more modern "new school" results are shaded (see Chapter 6)

With two regimes and two types of characterizations, there are four different characterizations. These four characterizations, together with their corresponding results, are summarized in Figure 4.5. Informally, we distinguish more traditional "old school" statistics and RMT results (in Theorem 5.1 on the asymptotic law of large numbers for SCMs, and in Theorem 5.4 on the non-asymptotic concentration of SCMs, both in the classical regime, as well as in Theorem 5.7 on the asymptotic Marčenko-Pastur distribution, in the proportional regime) from "new school" RMT results (in Theorem 6.5 and Theorem 6.7, establishing both asymptotic and non-asymptotic results in the proportional regime). The latter are more relevant for modern ML, and they are the main focus of our discussion in this monograph.

Chapter 5

Traditional RMT analysis of SCM eigenvalues

In this chapter, following the discussions on the classical versus proportional regime in Definition 4.23, we present "old school" results in Figure 4.5:

1. in the classical $n \gg p$ regime, both asymptotic and non-asymptotic characterizations of the sample covariance matrix (SCM) $\hat{\mathbf{C}}$ around the population covariance $\mathbf{C}$; and

2. in the proportional $n \sim p$ regime, *different* asymptotic behavior of the eigenvalue distribution of the SCM.

In more detail, by considering $n \to \infty$ with fixed p, the asymptotic behavior in the classical regime is via a law of large numbers (Theorem 5.1) in Chapter 5.1; and, for $n \gg p$ large but finite, the non-asymptotic behavior is via a matrix concentration result (Theorem 5.4) in Chapter 5.2. By considering the limiting behavior as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the asymptotic behavior in the proportional regime is via a traditional RMT result, the Marčenko-Pastur theorem (Theorem 5.7) in Chapter 5.3.

These results are well-known and should be intuitive to most of the ML audience (at least relative to some of the novel results we present), but we include them for completeness and for comparison with our main results (which we describe in Chapter 6 and subsequent chapters).

5.1 Classical regime: asymptotic behavior of SCM via LLNs

Let us start with $n \to \infty$ with p fixed. This corresponds to the asymptotic characterization in the classical regime in Definition 4.23. Assume that $\mathbf{C} = \mathbf{I}_p$, and consider each element of the SCM $\hat{\mathbf{C}}$. By the strong law of large numbers in Theorem 1.7, one has that

$$[\hat{\mathbf{C}}]_{ij} = \frac{1}{n} \sum_{l=1}^{n} [\mathbf{X}]_{il} [\mathbf{X}]_{jl} \xrightarrow{a.s.} \begin{cases} 1, & i = j; \\ 0, & i \neq j, \end{cases} \tag{5.1}$$

where $[\mathbf{X}]_{il}$ is the (i, l) entry of $\mathbf{X}$. Under a tail bound assumption on the entries of $\mathbf{X}$, the *entry-wise* convergence result in Equation (5.1) holds *uniformly* over all entries. That is,

$$\max_{1 \leq i,j \leq p} \left| [\hat{\mathbf{C}} - \mathbf{I}_p]_{ij} \right| \xrightarrow{a.s.} \delta_{ij}, \quad \text{as } n \to \infty.$$

Thus, the convergence in max norm

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_{\max} \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty,$$

holds, where $\|\mathbf{A}\|_{\max} \equiv \max_{ij} |\mathbf{A}_{ij}|$. Since $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq p\|\hat{\mathbf{C}} - \mathbf{I}_p\|_{\max}$ for any matrix $\hat{\mathbf{C}}$ of size p-by-p, it then follows that in spectral norm

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \xrightarrow{a.s.} 0, \quad \text{as } n \to \infty. \tag{5.2}$$

As such, by the strong law of large numbers, if p is fixed, then $\hat{\mathbf{C}} \to \mathbf{I}_p$ almost surely as $n \to \infty$. In this case, $\|\hat{\mathbf{C}} - \mathbf{I}_p\| \xrightarrow{a.s.} 0$ holds for *any* standard matrix norm, and in particular for the max and the spectral norm. *This result holds in the $n \gg p$ regime.* The following theorem makes this discussion more precise.

Theorem 5.1 (**Asymptotic Law of Large Numbers for SCMs**). *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*

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \to 0, \tag{5.3}$$

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

Proof of Theorem 5.1. By the definition of the SCM in Equation (4.46), one has

$$[\hat{\mathbf{C}}]_{ij} = \frac{1}{n}\sum_{l=1}^{n} [\mathbf{x}_l]_i [\mathbf{x}_l]_j, \tag{5.4}$$

where $[\mathbf{x}_l]_i$ the ith entry of the sub-gaussian random vector $\mathbf{x}_l \in \mathbb{R}^p$. As such, for $i = j$, the quantity $[\hat{\mathbf{C}}]_{ii} - 1 = \frac{1}{n}\sum_{l=1}^{n}([\mathbf{x}_l]_i^2 - 1)$ is the sum of n *independent* zero-mean sub-exponential random variables. (This is since any one-dimensional marginal of a sub-gaussian random vector is a sub-gaussian random variable, and the square of a sub-gaussian random variable is a sub-exponential random variable.) It then suffices to apply Bernstein's inequality for sub-exponential distribution (see, e.g., [36, Theorem 2.8.2]) to obtain

$$\mathbb{P}\left(\left|[\hat{\mathbf{C}}]_{ii} - 1\right| \geq t\right) \leq 2\exp\left(-C_1 n \min(t^2, C_2 t)\right),$$

for some constants $C_1, C_2 > 0$ that only depend on the sub-gaussian norm of the entries of $\mathbf{X}$. For $i \neq j$, one can similarly obtain

$$\mathbb{P}\left(\left|[\hat{\mathbf{C}}]_{ij}\right| \geq t\right) \leq 2\exp\left(-C_1 n \min(t^2, C_2 t)\right),$$

where we used the fact that the product of sub-gaussian random variables is a sub-exponential random variable, so that

$$\mathbb{P}\left(\left|[\hat{\mathbf{C}}]_{ij} - \delta_{ij}\right| \geq t\right) \leq 2\exp\left(-C_1 n \min(t^2, C_2 t)\right).$$

Taking the union bound, one obtains

$$\mathbb{P}\left(\max_{1\leq i,j,\leq p}\left|[\hat{\mathbf{C}} - \mathbf{I}_p]_{ij}\right| \geq t\right) \leq 2p^2\exp\left(-C_1 \min(t^2, C_2 t)\right).$$

Equivalently,

$$\mathbb{P}(\|\hat{\mathbf{C}} - \mathbf{I}_p\|_{\max} \geq t) \leq 2p^2\exp\left(-C_1 n \min(t^2, C_2 t)\right), \tag{5.5}$$

where we recall the definition of the max norm, $\|\mathbf{A}\|_{\max} \equiv \max_{ij}|\mathbf{A}_{ij}|$ of $\mathbf{A}$. Since $\|\mathbf{A}\| \leq p\|\mathbf{A}\|_{\max}$ for $\mathbf{A} \in \mathbb{R}^{p\times p}$, we further get

$$\mathbb{P}(\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \geq t) \leq \mathbb{P}(\|\hat{\mathbf{C}} - \mathbf{I}_p\|_{\max} \geq t/p) \leq 2\exp\left(-C_1 n \min(t^2, C_2 t)\right),$$

for some constants $C_1, C_2 > 0$ that only depend on the sub-gaussian norm of the entries of $\mathbf{X}$ *and* the dimension p. It then follows from the Borel–Cantelli lemma in Theorem A.1 of Appendix A that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \to 0$ almost surely as $n \to \infty$. See This concludes the proof of Theorem 5.1. □

Remark 5.2 (**Inverse of SCM**)**.** Theorem 5.1 states that, for fixed dimension p and in the limit of infinitely many samples as $n \to \infty$, the SCM $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ is close, in a spectral norm sense, to the population covariance $\mathbf{C} = \mathbf{I}_p$. In this $n \to \infty$ with p fixed regime, this in particular implies that the regularized SCM inverse $\mathbf{Q}(-\gamma) \equiv (\hat{\mathbf{C}} + \gamma\mathbf{I}_p)^{-1}$, should be close to the inverse population covariance $(\mathbf{C} + \gamma\mathbf{I}_p)^{-1}$ with the *same* regularization $\gamma > 0$. This is a consequence of the fact that

$$\|\mathbf{Q}(-\gamma) - (\mathbf{C} + \gamma\mathbf{I}_p)^{-1}\|_2 = \|\mathbf{Q}(-\gamma) \cdot (\mathbf{C} - \hat{\mathbf{C}}) \cdot (\mathbf{C} + \gamma\mathbf{I}_p)^{-1}\|_2 \leq \gamma^{-2}\|\mathbf{C} - \hat{\mathbf{C}}\|_2, \tag{5.6}$$

where we used the fact that $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ (known as the resolvent identity) for the equality, and that $\|(\mathbf{A} + \gamma\mathbf{I}_p)^{-1}\|_2 \leq \gamma^{-1}$ for all positive semi-definite $\mathbf{A}$ for the inequality. As well shall see below in Remark 6.11, *this conclusion is no longer valid in the proportional* $n \sim p \gg 1$ *regime.*

Remark 5.3 (**LLN and the classical versus proportional regime**)**.** Observe that the LLN in Theorem 5.1 is "parameterized" to hold only in the classical limit, not the proportional limit, and its proof fails in the limit of $n, p \to$ with $p/n \to c \in (0, \infty)$. There are many variants and extensions of the LLN; see, e.g., the non-asymptotic matrix concentration result in Theorem 5.4 below. Most – if not all – of them become vacuous when applied to the proportional regime where $n, p \to \infty$ and $p/n \to c \in (0, \infty)$. We will come back to this point in Remark 5.6 below, and we will clarify the reason behind this in Remark 5.8.

5.2 Classical regime: non-asymptotic behavior of SCM via matrix concentration

The asymptotic characterization of the SCM in Theorem 5.1 provides a precise statement in the classical limit with $n \to \infty$ with fixed p. We next use a spectral norm concentration bound on $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2$ to provide a more precise characterization of the SCM approximation $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \approx 0$ that is non-asymptotic, in the sense that it holds for *any* finite n, p.

Theorem 5.4 (**Non-asymptotic matrix concentration for SCMs, [36, 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*

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq C_1 \max(\delta, \delta^2), \quad \delta = C_2(\sqrt{p/n} + t/\sqrt{n}), \tag{5.7}$$

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

We reproduce the proof approach proposed in [36], which combines the Bernstein's concentration inequality with an ϵ-net argument.

Proof of Theorem 5.4. Using the ϵ-net argument (see, e.g., [36, Corollary 4.2.13]), one can find a 1/4-net $\mathfrak{N}$ of the unit sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ that has cardinality $|\mathfrak{N}| \leq 9^p$. The use of the ϵ-net technique allows one to well approximate the spectral norm via an evaluation over an ϵ-net $\mathfrak{N}$ of the unit sphere $\mathbb{S}^{p-1}$, rather than over the full unit sphere $\mathbb{S}^{p-1}$ itself. We refer the interested readers to [36, 37] for details. Then,

$$\left\|\hat{\mathbf{C}} - \mathbf{I}_p\right\|_2 = \sup_{\mathbf{v}\in\mathbb{S}^{p-1}} \left\|(\hat{\mathbf{C}} - \mathbf{I}_p)\mathbf{v}\right\|_2 \leq 2\max_{\mathbf{v}\in\mathfrak{N}} \left\|(\hat{\mathbf{C}} - \mathbf{I}_p)\mathbf{v}\right\|_2 = 2\max_{\mathbf{v}\in\mathfrak{N}} \left|\frac{1}{n}\|\mathbf{X}^\mathsf{T}\mathbf{v}\|_2^2 - 1\right|, \tag{5.8}$$

where one uses [36, Lemma 4.4.1] for the inequality and recalls the definition $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ for sub-gaussian $\mathbf{x}_i \in \mathbb{R}^p$ with $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. To complete the proof of Theorem 5.4, it then suffices to show, with the required probability, that

$$\max_{\mathbf{v}\in\mathfrak{N}} \left| \frac{1}{n}\|\mathbf{X}^\mathsf{T}\mathbf{v}\|^2 - 1 \right|_2 \leq \frac{\varepsilon}{2}, \tag{5.9}$$

with $\varepsilon = \max(\delta, \delta^2)/\sqrt{C_2}$. To that end, first note that for a *fixed* $\mathbf{v} \in \mathfrak{N}$ of unit norm $\|\mathbf{v}\|_2 = 1$, one has

$$\frac{1}{n}\|\mathbf{X}^\mathsf{T}\mathbf{v}\|_2^2 - 1 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\mathsf{T}\mathbf{v})^2 - 1. \tag{5.10}$$

By the sub-gaussianity of $\mathbf{x}_i$, one has that (i) $\mathbf{x}_i^\mathsf{T}\mathbf{v}$ is sub-gaussian (one-dimensional marginal of sub-gaussian random vector is sub-gaussian) with $\mathbb{E}[\mathbf{x}_i^\mathsf{T}\mathbf{v}] = 0$ and $\mathbb{E}[(\mathbf{x}_i^\mathsf{T}\mathbf{v})^2] = \mathbf{v}^\mathsf{T}\mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\mathsf{T}]\mathbf{v} = 1$; so that (ii) $(\mathbf{x}_i^\mathsf{T}\mathbf{v})^2$, as the square of sub-gaussian random variable, is thus sub-exponential; and therefore (iii) $\frac{1}{n}\|\mathbf{X}^\mathsf{T}\mathbf{v}\|_2^2 - 1$, as the sum of n independent zero-mean sub-exponential random variables, satisfies the following sub-exponential Bernstein inequality (see, for example, [36, Theorem 2.8.2]). For any $t \geq 0$, one has

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\|\mathbf{X}^\mathsf{T}\mathbf{v}\|_2^2 - 1\right| \geq \frac{\varepsilon}{2}\right) &\leq 2\exp\left(-C_1 n \min\left(C_2\varepsilon^2, \sqrt{C_2}\varepsilon\right)\right) = 2\exp\left(-C_1 n\delta^2\right) \\ &\leq 2\exp\left(-C_1 C_2^2(p + t^2)\right),\end{aligned}$$

where we used the fact that $\varepsilon = \max(\delta, \delta^2)/\sqrt{C_2}$ in the equality and the definition of δ in Equation (5.7) as well as the fact that $(a + b)^2 \geq a^2 + b^2$ for $a, b \geq 0$ in the last inequality.

It remains to apply the union bound to see

$$\begin{aligned}\mathbb{P}\left(\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq \varepsilon\right) &= \mathbb{P}\left(\max_{\mathbf{v}\in\mathfrak{N}}\left|\frac{1}{n}\|\mathbf{X}^\mathsf{T}\mathbf{v}\|_2^2 - 1\right| \leq \frac{\varepsilon}{2}\right) \\ &\leq 2\cdot 9^p \cdot \exp\left(-C_1 C_2^2(p + t^2)\right) \leq 2\exp(-t^2),\end{aligned}$$

by choosing the constant C_2 in Equation (5.7) large enough. This concludes the proof of Theorem 5.4. □

Remark 5.5 (**Derivation of Theorem 5.1 from Theorem 5.4**)**.** Instead of the simpler and more direct proof of Theorem 5.1 that we provided, one could alternatively prove Theorem 5.1 as a corollary of Theorem 5.4. To do so, take $t = \sqrt{2\ln n}$ to see that for $n \geq p + 2\ln n$, with probability at least $1 - 2n^{-2}$,

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq C_1 C_2(\sqrt{p} + \sqrt{2\ln n})/n, \tag{5.11}$$

with vanishing right-hand side as $n \to \infty$ with fixed p. It then follows from Borel–Cantelli lemma (in Theorem A.1 of Appendix A) that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \to 0$ almost surely at $n \to \infty$. This concludes the (alternate) proof of Theorem 5.1.

Remark 5.6 (**Matrix concentration and the classical versus proportional regime**)**.** The non-asymptotic result in Theorem 5.4 is practical, in that its holds for an arbitrary choice of n, p. Specifically, it should be compared to and contrasted with the asymptotic result in Theorem 5.1. Depending on the (classical versus proportional) regime in which one is operating, Theorem 5.4 conveys the following complementary messages.

1. **Classical regime.** Here, $n \gg p$. Let's say that $n \sim p^2$. In this case, one has, with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2$ is of order $O(n^{-1/4})$ and gets very small as n gets large. In this regime, where $n \gg p$, the matrix concentration in Theorem 5.4 conveys a similar intuition to the asymptotic LLN result in Theorem 5.1 and discussed in Remark 5.3.

2. **Proportional regime.** Here, n, p are both large, and in particular $n \sim p$. In this case, one has, with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2$ is of order $\sqrt{p/n} = O(1)$. In this regime, Theorem 5.4 is qualitatively different than Theorem 5.1: one can have, say, a vacuous 100% relative error, in the proportional limit of $n, p \to \infty$ with $p/n \to c \in (0, \infty)$.

Based on this discussion, the question then is the following: In the proportional regime, where n is *not much larger than* p, what precisely does the sample covariance $\hat{\mathbf{C}}$ "looks like"? For example, is it close, say, in a spectral sense, to its population counterpart $\mathbf{C} = \mathbf{I}_p$? Can we provide a more precise and quantitative description of, e.g., the maximum and minimum eigenvalues of the SCM $\hat{\mathbf{C}}$, or the "distribution" of the eigenvalues of $\mathbf{C}$ around the population eigenvalue 1? We discuss these topics next.

5.3 Proportional regime: eigenvalues via traditional RMT and Marčenko-Pastur

Here, we will show that, in the asymptotic proportional regime of $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the limiting eigenvalue distribution of $\hat{\mathbf{C}}$ takes a *precise* form, known as the Marčenko-Pastur distribution. This is a classical topic in traditional RMT. The Marčenko-Pastur distribution is a deterministic function whose shape is parameterized by the dimension ratio c and whose scale parametrized by a variance parameter σ^2; and it provides a more refined characterization of the eigenspectrum of $\hat{\mathbf{C}}$ (than is provided by Theorem 5.4). It is given in the following result, stated here in the case of sub-gaussian random vectors.[12] We provide in Remark 6.6 of Chapter 6.2 a proof of Theorem 5.7, as a consequence and corollary of our main Deterministic Equivalent for SCM resolvent in Theorem 6.5.

Theorem 5.7 (**Limiting spectral distribution for SCM: Marčenko-Pastur law, [21]**). *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 $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \sigma^2 \mathbf{I}_p$. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, with probability one, the empirical spectral measure $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ (as in Definition 2.20) converges weakly to a probability measure μ given explicitly by*

$$\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c \sigma^2 x} \sqrt{(x - \sigma^2 E_-)^+ (\sigma^2 E_+ - x)^+}\, dx, \tag{5.12}$$

where $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$. In particular, taking $\sigma^2 = 1$ in Equation (5.12), one obtains

$$\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ (E_+ - x)^+}\, dx, \tag{5.13}$$

which is known as the Marčenko-Pastur distribution.

The following remark on Theorem 5.7 should be compared to and contrasted with Remark 5.3 and 5.6.

Remark 5.8 (**Marčenko-Pastur law and the classical versus proportional regime**). The behavior described in Theorem 5.7 takes a very different form than the behavior of $\|\hat{\mathbf{C}} - \mathbf{C}\|_2 \approx 0$, as given by Theorem 5.1 and 5.4, in classical regime with $n \gg p$. Depending on the (classical versus proportional) regime of interest, Theorem 5.7 can lead to very different intuitions for the (eigenvalues of the) SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p \times p}$ composed of n samples.

[12] The sub-gaussian assumption here can be replaced by, e.g., random vectors having independent entries with a uniform bound on the moments of order k for some $k > 2$. Determining the minimalistic conditions for these RMT results to hold has been of long interest to mathematicians. We refer readers to [3, 5, 33] for more discussions.

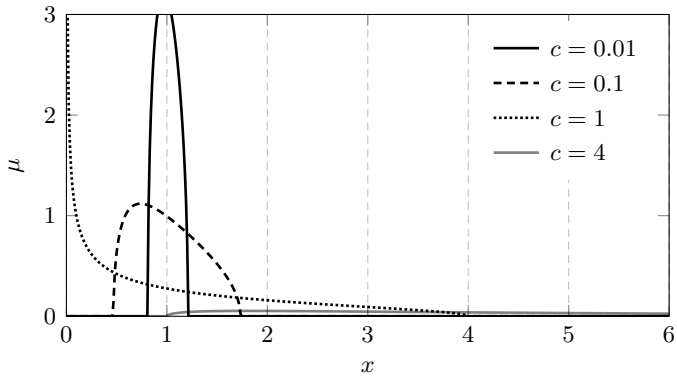


Figure 5.1: Marčenko-Pastur distribution for different values of c, for $\sigma^2 = 1$.

1. **Classical regime.** Here, $n \gg p$. Taking the dimension ratio $c = p/n \to 0$, the Marčenko-Pastur law in Equation (5.13) of Theorem 5.7 shrinks to δ_1, the Dirac measure at one. In this regime, it is in agreement with $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \simeq 0$ in Theorem 5.1 and 5.4.

2. **Proportional regime.** Here, $n \sim p \gg 1$. In this regime, it follows from the (true but vacuous) matrix concentration result in Theorem 5.4 that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(p/n) = O(1)$, and the (true but non-vacuous) result from Theorem 5.7 that, depending on the dimension ratio $c = p/n$, the eigenvalues of $\hat{\mathbf{C}}$ can be *very different* from unity. In particular, $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2$ is not vanishing small as $n, p \to \infty$, instead taking the form of the Marčenko-Pastur distribution given in Equation (5.13).

Remark 5.9 (**Precise behavior of the SCM eigenvalues**)**.** Theorem 5.7 provides access to the *averaged* amount of eigenvalues of $\hat{\mathbf{C}}$ lying within the interval $[1-\delta, 1+\delta]$, for $\delta \in (0,1)$. This can be seen by evaluating the following integral:

$$\mu([1-\delta, 1+\delta]) = \int_{1-\delta}^{1+\delta} \frac{1}{2\pi c x} \sqrt{\left(x - (1-\sqrt{c})^2\right)^+ \left((1+\sqrt{c})^2 - x\right)^+}\, dx. \tag{5.14}$$

Consider $\delta \ll 1$ in Equation (5.14). By Taylor-expanding the expression around $x = 1$ for (say) $c = p/n < 4$, one obtains

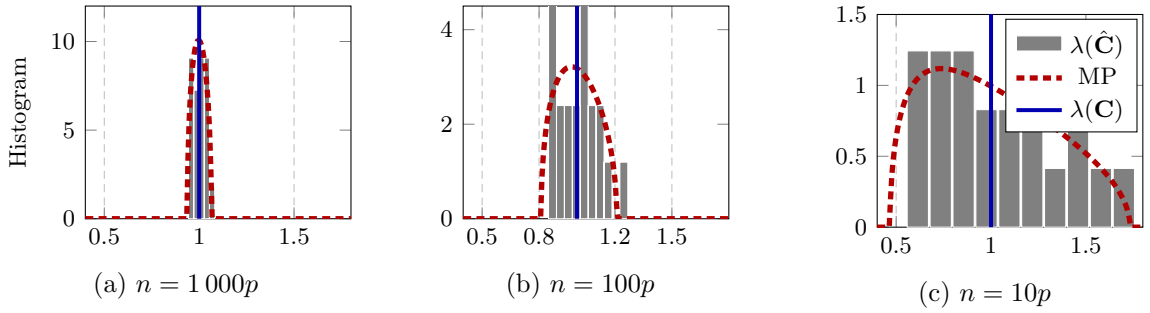
$$\begin{aligned}\mu([1-\delta, 1+\delta]) &= \int_{-\delta}^{\delta} \frac{1}{2\pi c(1+\varepsilon)} \sqrt{\left(1+\varepsilon - (1-\sqrt{c})^2\right)^+ \left((1+\sqrt{c})^2 - 1 - \varepsilon\right)^+}\, d\varepsilon \\ &= \frac{1}{2\pi c} \int_{-\delta}^{\delta} \left(\sqrt{4c - c^2} + O(\varepsilon)\right)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

Thus, in particular, for $p \approx 4n$ there is asymptotically *no* eigenvalue of $\hat{\mathbf{C}}$ close to one! This is in accordance with the shape of the limiting Marčenko-Pastur law with $c = 4$, displayed in Figure 5.1. More generally, one explicitly obtains from Equation (5.14) the limiting eigenvalue distribution of $\hat{\mathbf{C}} - \mathbf{I}_p$ as

$$(1 - c^{-1})^+ \delta_{-1}(x) + \frac{1}{2\pi c(x+1)} \sqrt{\left(x + 1 - (1-\sqrt{c})^2\right)^+ \left((1+\sqrt{c})^2 - x - 1\right)^+}\, dx, \tag{5.15}$$

where $\delta_{-1}(x)$ is the Dirac measure at $x = -1$. This provides access to the spectral norm[13] $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \simeq c + 2\sqrt{c}$ as well as more refined characterization such as the averaged amount of eigenvalues of $\hat{\mathbf{C}} - \mathbf{I}_p$ within a given interval of interest, as in Equation (5.14).

[13] Technically speaking, the limiting eigenvalue distribution given in Theorem 5.7 only characterizes the proportion of eigenvalues appearing within a given interval, and allows for an order of $o(p)$ eigenvalues that may "leak" from the interval. As a consequence, the Marčenko-Pastur law itself fails to assess the maximum or minimum eigenvalue of $\hat{\mathbf{C}}$ or $\hat{\mathbf{C}} - \mathbf{I}_p$, which needs additional efforts to characterize; see [2] and [6, Section 2.3.2]. Our conclusion on $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2$ here remains correct though.



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

Figure 5.2: **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 5.7, for $\mathbf{X}$ having standard Gaussian entries with $p = 20$ and different $n = 1\,000p, 100p, 10p$ from left to right.

Let us illustrate the results in Theorem 5.7 versus those in Theorem 5.1 and 5.4, as n and p vary. See Figure 5.2 and Figure 5.3. Here, we consider a *single realization* of the (Gaussian) random matrix $\mathbf{X}$.

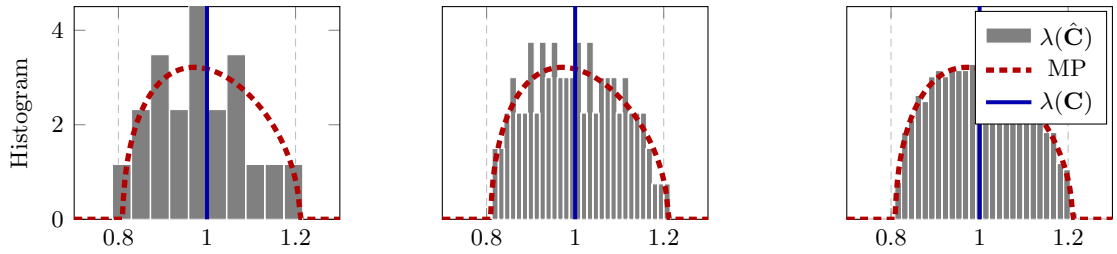
1. Figure 5.2 depicts the histogram of the eigenvalues of $\hat{\mathbf{C}}$ as we vary n (and thus $c = p/n$), for fixed p: for $p = 20$ and $n = 1\,000p, 100p, 10p$. We see a "transition" from the classical regime (Figure 5.2a with $n = 1\,000p \gg p$, in which case the random SCM $\hat{\mathbf{C}}$ strongly concentrates around $\mathbf{C} = \mathbf{I}_p$, as predicted by Theorem 5.1 and 5.4) to the proportional regime behavior (Figure 5.2c with $n = 10p \sim p$, in which case the eigenvalues of $\hat{\mathbf{C}}$ are more "spread out" and take a Marčenko-Pastur shape, given in Theorem 5.7).

2. Figure 5.3 illustrates what happens as we vary n and p together, for fixed $c = p/n$: for $p/n = c = 0.01$ with (in fact only moderately large) $p = 20, 100, 500$. This figure provides a "finite-dimensional" confirmation of the limiting Marčenko-Pastur law in Theorem 5.7. The eigenvalue histogram agrees with Marčenko-Pastur law, and this holds for *any* realization with n, p large, showing an *asymptotically deterministic* behavior of the behavior of the (distribution of the) eigenvalues of $\hat{\mathbf{C}}$. In particular, the Marčenko-Pastur law in Equation (5.13) demonstrates that the eigenvalues of $\hat{\mathbf{C}}$, instead of concentrating at $x = 1$, as the classical intuition would suggest, are spread from $E_- = (1-\sqrt{c})^2$ to $E_+ = (1+\sqrt{c})^2$. That is, they are on a range

$$(1 + \sqrt{c})^2 - (1 - \sqrt{c})^2 = 4\sqrt{c}. \tag{5.16}$$

   Observe that the convergence to the classical $n \gg p$ regime, as a function of the ratio $c = p/n$, is *not* very fast. In particular, even with $n = 100p$, one obtains an improved accuracy of $\pm 20\%$ by considering the proportional instead of the classical regime. This is numerically illustrated in Figure 5.3.

3. As a side remark, note that Figure 5.2b and 5.3a are two *independent realizations* of the case $p = 20$ and $n = 100p$ (with two different X-axis scalings, so we have intra-figure consistency). This provides an estimate of the sample-to-sample variability. In particular, the "shapes" of eigenvalue histograms remain random, differing from one realization to another (due to the intrinsic randomness in $\mathbf{X}$) in Figure 5.2b versus 5.3a, and they *cannot* be accurately described by either Theorem 5.1 or Theorem 5.7 (which are essentially *deterministic*). Also, while $n = 100p$ (with a sample size n that is 100 times the data dimension p) might seem "large enough" to be in the classical regime, we see that the eigenvalues of $\hat{\mathbf{C}}$ are very different from 1, being spread on the interval $[0.8, 1.2]$ (that diverges from 1 by a relative error of $\pm 20\%$).



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

Figure 5.3: **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.

By working in the special (yet realistic for modern ML) regime of $n \sim p \gg 1$, RMT analysis such as that in Theorem 5.7 allows for a more precise characterization of the (spectral) behavior of popular matrix models such as the sample covariance matrix. This is accomplished by describing how the full *distribution* of the eigenvalues of $\hat{\mathbf{C}}$ around unity,[14] and consequently the minimum and maximum eigenvalues, as well as the "proportion" of eigenvalues within *any* interval of interest.

Remark 5.10 (ESD of regularized inverse SCM). Similar to Remark 5.2, when the regularized inverse SCM $\mathbf{Q}(-\gamma) \equiv (\hat{\mathbf{C}} + \gamma \mathbf{I}_p)^{-1}$, $\gamma \geq 0$, is considered, Theorem 5.7 applies to assess the (limiting) eigenvalue distribution of $\mathbf{Q}(-\gamma)$ with the change of variable $x \mapsto 1/(x + \gamma)$ in Equation (5.13). Attention should be given to the inverse $1/(x + \gamma)$, which may *not* be well defined for $\gamma = 0$, depending on the sign of $c - 1$, for dimension ratio $c = p/n$. This is again an illustrative example of working in the proportional $n \sim p$ regime as opposed to its classical counterpart.

[14] From Theorem 5.4, this distribution of eigenvalues is only known to be of order $\sqrt{p/n} = O(1)$.

Chapter 6

SCM analysis beyond eigenvalues: a modern RMT approach

In this chapter, we discuss the "new school" results in Figure 4.5. In Chapter 5, we saw that, for the SCM $\hat{\mathbf{C}} \in \mathbb{R}^{p \times p}$ composed of n samples of dimension p, LLN and matrix concentration methods provide information about the eigenvalues of large random matrices in the classical $n \gg p$ regime (in Theorem 5.1 and Theorem 5.4), both asymptotically and non-asymptotically; and we also saw (in Theorem 5.7) how traditional RMT methods can be used to provide information, asymptotically, about eigenvalues in the proportional $n \sim p \gg 1$ regime. As we will see in this chapter, RMT methods are much more powerful.

They can be used to provide information, both asymptotically and non-asymptotically, for many other (eigenspectral) quantities, including those that depend on eigenvectors, of interest in modern ML and beyond. This is accomplished by analyzing more sophisticated spectral functionals of large random matrices that are of practical interest in a modern ML context. (These functionals go beyond the trace, or Stieltjes transform in **??**, which gives the limiting eigenvalue distribution in Theorem 5.7, to include functionals listed in **??** of **??**). To accomplish this, we must consider the *Deterministic Equivalent* approach to the SCM resolvent, which will be the major focus of this chapter.[15]

See Figure 6.1 for a high-level summary of the general approach. The figure compares the different objects of interest that can be analyzed with RMT: some with "old school" traditional RMT (that involve only the trace function, and for which only eigenvalues are considered) and "new school" modern RMT (that considers other eigenspectral functions, and that also considers eigenvectors) as well as the corresponding mathematical tools. The most important technical difference is the following:

> traditional "old school" RMT mostly focuses on eigenvalue distributions of large random matrices via a study of their Stieltjes transforms; while modern "new school" RMT works with the resolvent matrix directly, and as such is much more flexible.

The modern approach to RMT provides a simultaneous access to the behavior of large random matrices in the proportional regime via their *eigenspectral functionals*, as in **??**. This in particular includes spectral functionals involving both eigenvalues and eigenvectors that are of direct use in ML.

In Chapter 6.1, we first present the *Deterministic Equivalent for resolvent* framework, as a unified approach to evaluate the behavior of the resolvent of random matrices in the proportional regime. As an illustration of this approach, we provide in Chapter 6.2 and Chapter 6.3 asymptotic (in Theorem 6.5) and non-asymptotic (in Theorem 6.7) characterizations of the

[15] This chapter remains a work-in-progress, as the Stieltjes transform and the Deterministic Equivalent for resolvent approach have yet to be introduced. Nevertheless, we include it here to complete the broader discussion in Part II on the four ways of characterizing sample covariance matrices.

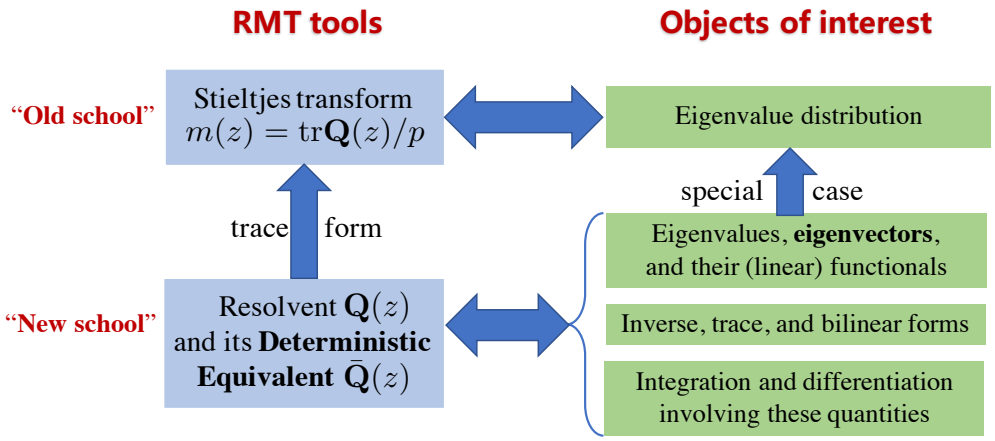


Figure 6.1: Different objects of interest and their corresponding technical tools for "old school" traditional RMT and "new school" modern RMT.

Deterministic Equivalent for the resolvent $\mathbf{Q}_{\hat{\mathbf{C}}}(z)$ of the SCM $\hat{\mathbf{C}}$, respectively. These characterizations provide analogues of Theorem 5.1 and Theorem 5.4 to the proportional regime, and they complete the proposed four-way taxonomy given in Figure 4.5. Some discussions on Theorem 6.5 and 6.7 are placed after these results, so as to connect the results in Theorems 6.5 and 6.7 to those in Theorems 5.1 and 5.4.

6.1 Deterministic Equivalents: from vectors to resolvent matrices

Here, we introduce the modern *Deterministic Equivalent for resolvent* approach to characterize the statistical behavior of eigenvalues and eigenvectors of a matrix $\mathbf{X}$, for $\mathbf{X}$ drawn from some "generic" random matrix model. We will present the asymptotic and non-asymptotic results for $\mathbf{X}$ following a sample covariance model (in Chapter 6.2 and Chapter 6.3, respectively).

Limitations of the Stieltjes transform and traditional "old school" RMT. In traditional "old school" RMT, the main focus is typically on the ESD of $\mathbf{X} \in \mathbb{R}^{p\times p}$. That is, the interest is in eigenvalues, in the characterization of a certain limit of the random spectral measure $\mu_{\mathbf{X}}$ (see again Definition 2.20) of $\mathbf{X}$, as the size p of $\mathbf{X}$ increases to infinity. The well-known Marčenko-Pastur law in Theorem 5.7 provides a canonical example of this approach. For this purpose (recall **??**), a natural approach is to study the *random Stieltjes transform* $m_{\mu_{\mathbf{X}}}(z)$ and show that it admits a limit (in probability or almost surely) $m(z)$ as $p \to \infty$. While leading to a large body of results in RMT, this approach has several strong limitations, in particular for modern ML applications. The main limitations include:

1. it supposes that such a limit exists, therefore restricting the study to very regular random matrix models for $\mathbf{X}$; and
2. it only quantifies the functional $\frac{1}{p}\mathrm{tr}\,\mathbf{Q}_{\mathbf{X}}(z)$ (through the Stieltjes transform), thereby discarding all eigenspace information about $\mathbf{X}$ carried in the resolvent matrix $\mathbf{Q}_{\mathbf{X}}$; and
3. it focuses only on the limiting behavior as $p \to \infty$ and in general fails to say any about *large but finite* p, which is of core interest to ML.

Deterministic Equivalents. To avoid these limitations of traditional RMT, and to provide "finite-dimensional" or non-asymptotic characterization of the quantities of interest in ML appli-

cations, modern "new school" RMT focuses instead on the notion of *Deterministic Equivalents*. Here is the basic definition.

High-dimensional Deterministic Equivalent

Definition 6.1 (**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*

$$\left|\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\bar{\mathbf{Q}})\right| \leq \varepsilon_1(p), \quad \left|\mathbf{a}^\mathsf{T}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{b}\right| \leq \varepsilon_2(p), \tag{6.1}$$

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*

$$\mathbf{Q} \overset{\varepsilon_1,\varepsilon_2,\delta}{\longleftrightarrow} \bar{\mathbf{Q}}, \text{ or simply } \mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}. \tag{6.2}$$

The Deterministic Equivalent relation $\mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}$ denotes the fact that $\bar{\mathbf{Q}}$, being a deterministic matrix, can be used as a "proxy" in the study of the large random resolvent matrix $\mathbf{Q}$, as long as its trace, bilinear forms, and other matrix functionals (as well as their differentiations and integrations as listed in **??**) are considered. The "accuracy" of this approximation for scalar observation is described by the error functions[16] $\varepsilon_1, \varepsilon_2$ and the failure probability δ. In particular, the Deterministic Equivalent in Definition 6.1 is a special case of the High-dimensional Equivalent in Definition 1.1, when the (random) matrix of interest if the resolvent $\mathbf{Q}(z)$, in the absence of entry-wise non-linearity ϕ, and for trace and bilinear observations $f(\mathbf{Q}) = \operatorname{tr}(\mathbf{X})/p$ and $f(\mathbf{X}) = \mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$ that are both 1-Lipschitz for $\mathbf{A} \in \mathbb{R}^{p\times p}, \mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of unit norm.

Remark 6.2 (**Asymptotic versus non-asymptotic aspects of of Deterministic Equivalents**). Definition 6.1 is *non-asymptotic*, in the sense that it holds for any value of p. As such, it can be used as a basis to provide both non-asymptotic as well as asymptotic results. By considering the *asymptotic* setting, as $p \to \infty$, one has

- by definition of convergence in probability, that $\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\bar{\mathbf{Q}}) \to 0$, $\mathbf{a}^\mathsf{T}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{b} \to 0$ in probability as $p \to \infty$; and
- if, in addition, the failure probability satisfies $\delta(p) = O(p^{-\ell})$ for some $\ell > 1$, then by the Borel–Cantelli lemma (see Theorem A.1 of Appendix A) $\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q}-\bar{\mathbf{Q}}) \to 0$, $\mathbf{a}^\mathsf{T}(\mathbf{Q}-\bar{\mathbf{Q}})\mathbf{b} \to 0$ almost surely as $p \to \infty$.

This "general recipe" has been used in Theorem 5.1 to obtain asymptotic (convergence-type) results from non-asymptotic results for SCM, and will be exploited in the reminder of the monograph.

Remark 6.3 (**Scalar observation function**). The notion of Deterministic Equivalents focuses on a scalar observation of the random matrix (recall Chapter 1.5 in Chapter 1). It does so by describing the *concentration behavior* of the random matrix via an observation map $f\colon \mathbb{R}^{p\times p} \to \mathbb{R}$ (as in Definition 1.17 for vectors). In Definition 6.1 above, there are two observation functions, one for each of the two expressions in Equation (6.1), and they take the form

$$f(\mathbf{X}) = \frac{1}{p}\operatorname{tr}(\mathbf{A}\mathbf{X}) \text{ and } f(\mathbf{X}) = \mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{b}. \tag{6.3}$$

[16] Here we distinguish the two "rates of convergence" $\varepsilon_1(\cdot), \varepsilon_2(\cdot)$ for trace and bilinear form, since they have been extensively investigated in the random matrix literature and are observed to take rather different forms. For instance, one commonly has $\varepsilon_1(p) \simeq p^{-1}$ while $\varepsilon_2(p) \simeq p^{-1/2}$ as $p \to \infty$; see also [6].

Remark 6.4 (**Deriving Deterministic Equivalents**)**.** Mathematically, the derivation of a Deterministic Equivalent is generally accomplished via the following two steps:

1. **Computing or approximating the expectation of the random matrix Q.** For the scalar random variable of interest $f(\mathbf{Q})$ for $\mathbf{Q} \in \mathbb{R}^{p\times p}$, the first (and often most natural) *deterministic* quantity to describe its behavior is the expectation $\mathbb{E}[f(\mathbf{Q})]$.

   - In the case of linear or bilinear functional $f(\cdot)$, as in Definition 6.1, this is equal to $f(\mathbb{E}[\mathbf{Q}])$.
   - In the case where $\mathbb{E}[\mathbf{Q}]$ is not easily accessible, one may resort to approximating it using some deterministic matrix $\bar{\mathbf{Q}}$, rather than directly computing it (e.g., in the sense that $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \varepsilon(p)$ for some function $\varepsilon(\cdot)$ that vanishes as p grow large).

   In this sense, a Deterministic Equivalent for a random matrix $\mathbf{Q}$ is not necessarily unique. See Remark 6.11 below for a concrete example of this non-uniqueness.

2. **Establishing the concentration of the random observation $f(\mathbf{Q})$ around the deterministic $f(\bar{\mathbf{Q}})$.** This step often involves concentration inequalities of the form
$$\mathbb{P}(|f(\mathbf{Q}) - f(\bar{\mathbf{Q}})| \geq t) \leq \delta(p,t) \tag{6.4}$$
for some function $\delta(p,t)$ that decreases sufficiently fast as p grows large. This can be achieved, e.g., by bounding sequentially, in a probabilistic sense and as in Chapter 1.3 for scalar observations of the rando vectors, the differences $f(\mathbf{Q}) - f(\mathbb{E}[\mathbf{Q}])$ and $f(\mathbb{E}[\mathbf{Q}]) - f(\bar{\mathbf{Q}})$. (The latter uses the fact that the two *deterministic* matrices $\mathbb{E}[\mathbf{Q}]$ and $\bar{\mathbf{Q}}$ are close, in a spectral norm sense, as established in the first step.)

In Chapter 6.2 and 6.3 below, we will use this Deterministic Equivalent approach to establish asymptotic as well as non-asymptotic characterizations for the resolvent $\mathbf{Q}_{\hat{\mathbf{C}}}(z)$ for the SCM, respectively.

6.2 Asymptotic Deterministic Equivalents for SCM resolvents

Here, we illustrate the use of the proposed Deterministic Equivalent framework, by providing an asymptotic characterization of the random sample covariance resolvent

$$\mathbf{Q}(z) \equiv \mathbf{Q}_{\hat{\mathbf{C}}}(z) = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}, \tag{6.5}$$

for $\mathbf{X} \in \mathbb{R}^{p\times n}$ having i.i.d. "normalized" sub-gaussian entries with zero mean and unit variance. This result can be used to derive the popular Marčenko-Pastur law in Theorem 5.7, and it is a special case of our Linear Master Theorem, **??**, used to assess the three linear ML models in **??**.

Theorem 6.5 (**An asymptotic Deterministic Equivalent for resolvent, [6, 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 (sequence of) deterministic matrix $\bar{\mathbf{Q}}(z)$ is a Deterministic Equivalent of the (sequence of) random resolvent matrix $\mathbf{Q}(z)$ as in Definition 6.1 with*

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{6.6}$$

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

$$czm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{6.7}$$

One could prove Theorem 6.5 directly, but we prefer to prove it as a consequence of Theorem 6.7 (below), analogous to what we described in Remark 5.5 for the classical regime.

Proof of Theorem 6.5. The proof of Theorem 6.5 follows from that of Theorem 6.7 and concentration results on the trace and bilinear forms of the type $\frac{1}{p}\operatorname{tr}\mathbf{A}\mathbf{Q}$ and $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$ around their expectations. Precisely, it follows from the proof of Theorem 6.7 that for $\mathbf{A} \in \mathbb{R}^{p\times p}$ of unit norm, one has, as $n, p \to \infty$ that:

1. $\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 = O(n^{-1/2})$; and
2. $\mathbb{E}[\Big(\frac{1}{p}\operatorname{tr}\mathbf{A}(\mathbf{Q} - \mathbb{E}[\mathbf{Q}])\Big)^4] = O(n^{-2})$.

As such, by Markov's inequality (i.e., $\mathbb{P}(|X| \geq t) \leq \mathbb{E}[|X|^k]/t^k$) and the Borel–Cantelli lemma (i.e., $\mathbb{P}(|X| \geq t) = O(n^{-\ell})$ for $\ell > 1$ and all $t > 0$ implies $X_n \to 0$ almost surely as $n \to \infty$), it follows that

$$\frac{1}{p}\operatorname{tr}\mathbf{A}\mathbf{Q} - \frac{1}{p}\operatorname{tr}\mathbf{A}\mathbb{E}[\mathbf{Q}] \to 0 \tag{6.8}$$

almost surely as $n, p \to \infty$. Thus, the conclusion $\operatorname{tr}\mathbf{A}(\mathbf{Q} - \bar{\mathbf{Q}})/p \to 0$ follows almost surely. A similar procedure can be performed on the bilinear form $\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{b}$. This concludes the proof of Theorem 6.5 for all real $z < 0$. For complex $z \in \mathbb{C}$ not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, since both $\mathbf{Q}(z)$ and $\bar{\mathbf{Q}}(z)$ in the theorem statement are complex analytic functions on their domain of definition (see, e.g., [17]), by Vitali's convergence theorem, Theorem A.2, the convergence results for all $z < 0$ extend to $z \in \mathbb{C}$. This concludes the proof of Theorem 6.5. □

The function $m(z)$ in Theorem 6.5 is the *Stieltjes transform* of limiting spectral/eigenvalue distribution of the SCM. Thus, not surprisingly, one is able to retrieve the Marčenko-Pastur law in Theorem 5.7 from Theorem 6.5, as described in the following remark.

Remark 6.6 (**Derivation of Theorem 5.7, the Marčenko-Pastur law, from Theorem 6.5**). The equation of $m(z)$ in Equation (6.7) is quadratic, and thus it has two solutions defined via the (two values of) complex square root. That is, for $z = \rho e^{\imath\theta}$ for radius $\rho \geq 0$ and angle $\theta \in [0, 2\pi)$, we have $\sqrt{z} \in \{\pm\sqrt{\rho}e^{\imath\theta/2}\}$ and therefore

$$m(z) = \frac{1 - c - z}{2cz} + \frac{\sqrt{((1+\sqrt{c})^2 - z)\,((1-\sqrt{c})^2 - z)}}{2cz}. \tag{6.9}$$

Among these, only one satisfies the relation $\Im[z] \cdot \Im[m(z)] > 0$ as a "valid" Stieltjes transform of a probability measure, per its definition in **??**. By **??**, one obtains that $m(z)$ is the Stieltjes transform of μ: with "continuous" part of the form $\frac{\sqrt{(E_+ - x)^+(x - E_-)^+}}{2c\pi x}$ for $E_\pm = (1 \pm \sqrt{c})^2$ and

$(x)^+ = \max(x, 0)$ (since the term under the square root is only non-negative for $x \in [E_-, E_+]$); and with a discontinuity at zero with weight equal to

$$\mu(\{0\}) = -\lim_{y \downarrow 0} \imath y m(\imath y) = \frac{c-1}{2c} \pm \operatorname{sign}(c-1)\frac{c-1}{2c}, \tag{6.10}$$

and therefore a mass $1 - 1/c$ at zero if and only if $c > 1$.

6.3 Non-asymptotic Deterministic Equivalents for SCM resolvents

Here, we focus on the non-asymptotic characterization of the random resolvent matrix $\mathbf{Q}(z)$ of a SCM. We provide, for $z < 0$, a spectral norm error bound of $\|\mathbb{E}[\mathbf{Q}(z)] - \bar{\mathbf{Q}}(z)\|_2$ that depends *explicitly* on the dimension n, p, under the same statistical assumptions as in Theorem 6.5. Our main theorem, Theorem 6.7 below, is indeed a non-asymptotic version of the result in Theorem 6.5, and it allows one to have a precise control on, e.g., the approximation error of using the Deterministic Equivalent $\bar{\mathbf{Q}}(z)$ in place of the expected resolvent $\mathbb{E}[\mathbf{Q}(z)]$. This is of direct interest in ML, where non-asymptotic-type results are strongly desired.

Theorem 6.7 (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 real $z < 0$. Then, there exist 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*

$$\left\|\mathbb{E}[\mathbf{Q}(z)] - \bar{\mathbf{Q}}(z)\right\|_2 \leq \frac{C_2}{\varepsilon} \cdot n^{-\frac{1}{2}}, \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{6.11}$$

for $m(z)$ the unique positive *solution to the Marčenko-Pastur equation $czm^2(z) - (1 - c - z)m(z) + 1 = 0, c = p/n$ as in Equation* (6.7).

The proof of Theorem 6.7 provide a general recipe to get non-asymptotic Deterministic Equivalents for common random matrix models, and it is given in details as follows.

Proof of Theorem 6.7. Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the i^{th} 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 i^{th} column $\mathbf{x}_i$ (so that $\mathbf{X}_{-i}$ is in particular *independent* of $\mathbf{x}_i$). 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

$$\mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}_{-i}\mathbf{X}_{-i}^\mathsf{T} + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T} - z\mathbf{I}_p\right)^{-1} = \left(\mathbf{Q}_{-i}^{-1}(z) + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)^{-1}. \tag{6.12}$$

First note that by definition,

$$\bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p = \left(\frac{1}{1 + cm(z)} - z\right)^{-1}\mathbf{I}_p, \tag{6.13}$$

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

$$\frac{1}{1 + cm(z)}\|\bar{\mathbf{Q}}\|_2 \leq 1. \tag{6.14}$$

Similarly, one has

$$\|\mathbf{Q}(z)\|_2 \leq \frac{1}{|z|}, \quad \left\|\mathbf{Q}(z)\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\right\|_2 \leq 1, \quad \left\|\mathbf{Q}(z)\frac{1}{\sqrt{n}}\mathbf{X}\right\|_2 = \sqrt{\left\|\mathbf{Q}(z)\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{Q}(z)\right\|_2} \leq \frac{1}{\sqrt{|z|}}. \tag{6.15}$$

In the remainder of the proof, we will, for notational simplicity, drop the argument z and simply write $\mathbf{Q} = \mathbf{Q}(z)$, $\mathbf{Q}_{-i} = \mathbf{Q}_{-i}(z)$, and $\bar{\mathbf{Q}} = \bar{\mathbf{Q}}(z)$.

It follows from the resolvent identity, Lemma A.7, that

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

where we applied Woodbury identity, Lemma A.5, to obtain the last equality.

Further write

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

where we have introduced

$$d_i = \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i - cm(z), \tag{6.16}$$

and used again Woodbury identity to write $\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i} = \mathbf{Q}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ in the first equality, as well as the fact that the law of the random matrix $d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$ is *independent* of the index i in the second equality. Since $\mathbb{E}[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}] = \mathbb{E}[\mathbf{Q}_{-i}]$ by independence and the law of $\mathbf{Q}_{-i}$ is independent of the index i, this can be expressed as

$$\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}] = \left(\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\right)\frac{\bar{\mathbf{Q}}}{1+cm(z)} + \frac{\mathbb{E}\left[d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\bar{\mathbf{Q}}}{1+cm(z)}. \tag{6.17}$$

As such, to bound the spectral norm $\|\mathbb{E}[\mathbf{Q} - \bar{\mathbf{Q}}]\|_2$, it suffices to bound the following two quantities

$$T_1 = \|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\|_2, \quad T_2 = \left\|\mathbb{E}\left[d_i\mathbf{Q}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right]\right\|_2, \tag{6.18}$$

and then use the fact that $\|\mathbf{A}\mathbf{B}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{B}\|_2$, together with the bound in Equation (6.13).

For the first term T_1, it follows again from Woodbury identity that

$$0 \preceq \mathbb{E}[\mathbf{Q}_{-i} - \mathbf{Q}] = \mathbb{E}\left[\frac{\mathbf{Q}_{-i}\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}\right] \preceq \frac{1}{n}\mathbb{E}[\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}] = \frac{1}{n}\mathbb{E}\left[\mathbf{Q}_{-i}^2\right] \tag{6.19}$$

where we used the fact that $1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i \geq 1$, so that by Equation (6.15) one has $\|\mathbf{Q}_{-i}\| \leq |z|^{-1}$, and therefore

$$T_1 = \|\mathbb{E}[\mathbf{Q} - \mathbf{Q}_{-i}]\|_2 = O(n^{-1}). \tag{6.20}$$

We now move on to bound the second quantity T_2 as defined in Equation (6.18).

It follows from the definition of the spectral norm that

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

where we have applied the Cauchy-Schwarz inequality twice.

We first treat 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

$$\|\mathbf{Q}_{-i}\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}\|_2 = \mathbf{u}^\mathsf{T}\mathbf{Q}_{-i}^2\mathbf{u} \leq |z|^{-2}, \tag{6.21}$$

for $\|\mathbf{u}\| = 1$ according to Equation (6.15). As such, it follows from the Hanson–Wright inequality, Theorem 1.22, that there exists $C, C' > 0$ such that

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

where we first consider the expectation with respect to $\mathbf{x}_i$ and then that with respect to $\mathbf{Q}_{-i}$. This allows us to conclude that $T_{2,2} = O(1)$. And we can analogously conclude that $T_{2,3} = O(1)$.

We thus have

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq T_1 + T_2 \leq C_1 n^{-1} + C_2\sqrt{\mathbb{E}[d_i^2]}, \tag{6.22}$$

for some universal constants C_1, C_2 and $d_i \equiv \frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i - cm(z)$ as defined in Equation (6.16).

Now, note that

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

so that

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

where the expectation of the cross terms vanishes due to the independence between $\mathbf{Q}_{-i}$ and $\mathbf{x}_i$.

For the term D_1, it follows from the same line of arguments as the term $T_{2,2}$ above that $D_1 \leq Cn^{-2}$ for some constant $C > 0$.

For the term D_2, which characterizes the concentration property of the resolvent trace $\operatorname{tr} \mathbf{Q}_{-i}$, it can be bounded via a martingale difference argument using the Burkholder inequality, Lemma A.11.

For the sake of further use (e.g., in the proof of Theorem 6.5), we will prove a slightly more general result on $\mathbb{E}[(\operatorname{tr} \mathbf{Q}_{-i} - \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}])^2]$. First note that by Lemma A.6 we may freely replace $\mathbf{Q}_{-i}$ with $\mathbf{Q}$ without altering the desired bound, and that we may generalize the bound to $\mathbb{E}[(\operatorname{tr} \mathbf{AQ} - \operatorname{tr} \mathbb{E}[\mathbf{AQ}])^2]$ for any deterministic matrix $\mathbf{A}$ of unit spectral norm, that is, such that $\|\mathbf{A}\|_2 = 1$.

Specifically, under the notation of Lemma A.11, observe that we may write

$$\begin{aligned}\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbb{E}\mathbf{Q} - \mathbf{Q}) &= \sum_{i=1}^{n} \left(\mathbb{E}_i \left[\frac{1}{n} \operatorname{tr} \mathbf{AQ} \right] - \mathbb{E}_{i-1} \left[\frac{1}{n} \operatorname{tr} \mathbf{AQ} \right] \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\mathbb{E}_i - \mathbb{E}_{i-1}) \left[\operatorname{tr}(\mathbf{AQ} - \mathbf{AQ}_{-i}) \right],\end{aligned}$$

(since $\mathbb{E}_i[\operatorname{tr} \mathbf{AQ}_{-i}] = \mathbb{E}_{i-1}[\operatorname{tr} \mathbf{AQ}_{-i}]$) for $\mathcal{F}_i$ the σ-field generating the columns $\mathbf{x}_{i+1}, \ldots, \mathbf{x}_n$ of $\mathbf{X}$ and with the convention $\mathbb{E}_0[f(\mathbf{X})] = f(\mathbf{X})$. This forms a martingale difference sequence so that we fall under the scope of Burkholder inequality. Now, from the identity $\mathbf{Q} = \mathbf{Q}_{-i} - \frac{1}{n} \frac{\mathbf{Q}_{-i}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}$ (by Lemma A.5), we have that

$$\begin{aligned}\left| (\mathbb{E}_i - \mathbb{E}_{i-1}) \left[\frac{1}{n} \operatorname{tr}(\mathbf{AQ}_{-i} - \mathbf{AQ}) \right] \right| &= \left| (\mathbb{E}_i - \mathbb{E}_{i-1}) \frac{1}{n} \frac{\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{A}\mathbf{Q}_{-i}\mathbf{x}_i}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i} \right| \\ &\leq \frac{1}{n|z|} \cdot \left| (\mathbb{E}_i - \mathbb{E}_{i-1}) \frac{\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}\mathbf{x}_i} \right| \\ &\leq \frac{2}{n|z|}.\end{aligned}$$

As a consequence, it follows from Lemma A.11 that

$$\mathbb{E}\left[\left(\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \mathbb{E}\mathbf{Q}) \right)^2 \right] \leq Cn^{-1} \text{ and } \mathbb{E}\left[\left(\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \mathbb{E}\mathbf{Q}) \right)^4 \right] \leq Cn^{-2}, \tag{6.23}$$

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

Having obtained the above bounds on both D_1 and D_2, we can thus conclude that

$$\mathbb{E}[d_i^2] \leq 2(D_1 + D_2) + 2 \left(\frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] - cm(z) \right)^2 \leq Cn^{-1} + 2 \left(\frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}_{-i}] - cm(z) \right)^2, \tag{6.24}$$

for some universal constant $C > 0$. Therefore, from Equation (6.22) and Lemma A.6, it follows that

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq C_1 n^{-\frac{1}{2}} + C_2 \left| \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}] - cm(z) \right|. \tag{6.25}$$

Further note from Equation (6.13) that $\frac{1}{n} \operatorname{tr} \bar{\mathbf{Q}} = \frac{p}{n} m(z) = cm(z)$, so that

$$\left| \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}] - cm(z) \right| \leq \frac{p}{n} \|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \frac{p}{n} \left(C_1 n^{-\frac{1}{2}} + C_2 \left| \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}] - cm(z) \right| \right), \tag{6.26}$$

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

$$\left| \frac{1}{n} \operatorname{tr} \mathbb{E}[\mathbf{Q}] - cm(z) \right| \leq \frac{C_1}{\varepsilon} \cdot n^{-\frac{1}{2}}, \tag{6.27}$$

and thus

$$\|\mathbb{E}[\mathbf{Q}] - \bar{\mathbf{Q}}\|_2 \leq \frac{C}{\varepsilon} \cdot n^{-\frac{1}{2}}, \tag{6.28}$$

for some universal constant $C > 0$. This concludes the proof of Theorem 6.7. □

Theorem 6.5 and 6.7 extend, in an asymptotic and non-asymptotic manner, respectively, the LLN and matrix concentration results in Theorem 5.1 and 5.4, by providing precise characterization of the *expectation* of the random resolvent $\mathbf{Q}_{\hat{\mathbf{C}}}(z)$ of the SCM $\hat{\mathbf{C}}$. This characterization is technically challenging, due to the *nonlinear* matrix inverse in $\mathbf{Q}_{\hat{\mathbf{C}}}(z) = (\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1}$; but it is of great significance in the proportional $n \sim p$ regime.

A few remarks on Theorem 6.5 and 6.7 are in order.

Remark 6.8 (**Extension of Theorem 6.7 to** $z = 0$)**.** Theorem 6.7 is stated for any negative $z < 0$. The condition $z < 0$ is crucial in the proof presented above since it allows for a direct control 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 results in Theorem 6.7 to, say, an inverse SCM of the type

$$\mathbf{Q}(z = 0) = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \right)^{-1},$$

with $z = 0$, one first needs to ensure the inverse is properly defined for sub-gaussian $\mathbf{X}$ and for a specific choice of p, n. An improved bound can be obtained by considering the concentration of the sample covariance $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ around its expectation. For instance, it follows from Theorem 5.4 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

$$\left\| \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \mathbf{I}_p \right\|_2 \leq \frac{1}{2}, \tag{6.29}$$

and therefore $\|\mathbf{Q}(z)\|_2 \leq \frac{1}{1/2 - z} \leq 2$ for any $z \leq 0$. This allows for a control of the spectral norm $\|\mathbf{Q}(z)\| \leq 2$ that is independent of $z \leq 0$ and holds with probability at least $1 - \delta$. Within RandNLA, this strategy has been adopted [10], where results similar to Theorem 6.7 have been proved, by replacing the expectation $\mathbb{E}[\mathbf{Q}(z = 0)]$ with a conditional expectation $\mathbb{E}[\mathbf{Q}(z = 0) \mid \mathcal{E}]$ on an event $\mathcal{E}$ that holds with probability at least $1 - \delta$ and ensures the inverse $(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})^{-1}$ is well defined as in Equation (6.29).

Remark 6.9 (**Inversion bias**)**.** Continuing with Remark 6.8, the fact that for a symmetric and non-singular random matrix $\mathbf{X}$, one in general has that

$$\mathbb{E}[\mathbf{X}^{-1}] \neq (\mathbb{E}[\mathbf{X}])^{-1} \tag{6.30}$$

is referred to as the *inversion bias* [10]. The inversion bias has direct consequences for ML. As an instance, the authors of [10] identified the difference between $(\mathbb{E}[\hat{\mathbf{C}}])^{-1}$ and $\mathbb{E}[\hat{\mathbf{C}}^{-1}]$ in the context where $\hat{\mathbf{C}}$ is a sketched estimate of some covariance matrix; and they show how this difference impacts the performance of statistical inference and the convergence rate of distributed optimization. Also, although we have proven it here only for $\mathbf{X}$ having sub-gaussian entries, results of the type provided in Theorem 6.7 have been extended to a wide range of random matrices of interest in RandNLA [8, 11]. This includes so-called LEverage Score Sparsified (LESS) sketching matrices that have numerous numerical advantages, e.g., in stochastic optimization [9, 10].

Remark 6.10 (**Theorem 6.5 and 6.7 as extensions of Theorem 5.1 and 5.4**)**.** Theorems 6.5 and 6.7 provide characterizations of the SCM in the proportional, and should be compared and contrasted to Theorems 5.1 and 5.4 in the classical regime. Precisely, depending on the dimension ratio p/n, we have the following dual observation:

1. **Classical regime.** 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

$$(\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1} \simeq (\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1} = (1-z)^{-1}\mathbf{I}_p = \bar{\mathbf{Q}}(z). \tag{6.31}$$

2. **Proportional regime.** In the "harder" proportional regime, for $n \sim p$ with $p/n \to c \in (0, \infty)$, one has instead

$$\bar{\mathbf{Q}}(z) \simeq \mathbb{E}[\mathbf{Q}(z)] \equiv \mathbb{E}[(\hat{\mathbf{C}} - z\mathbf{I}_p)^{-1}] \not\simeq (\mathbb{E}[\hat{\mathbf{C}}] - z\mathbf{I}_p)^{-1}. \tag{6.32}$$

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

Equation (6.32) in the proportional regime is *not* surprising since the matrix inverse is *not* a linear operator, and so one can *not* swap the expectation and the inverse.[17] This observation on the random resolvent matrix and its Deterministic Equivalent explains the different between the spectral behaviors of $\hat{\mathbf{C}}$ in Theorem 5.1 and 5.4 for $n \gg p$ and in Theorem 5.7 and 6.7, for $n \sim p$ with $p/n \to c \in (0, \infty)$. The former is indeed as special case of the latter. It holds due to the convergence $\hat{\mathbf{C}} \to \mathbf{C} = \mathbf{I}_p$ that gets rid of the intrinsic non-linearity (due to inverse) in the evaluation of eigenvalues and eigenvectors.

Non-uniqueness of Deterministic Equivalents. We have said that, for a given random matrix model of interest, Deterministic Equivalents are not necessarily unique. For example, it suffices that they approximate the expectation $\mathbb{E}[\mathbf{Q}(z)]$ up to small error terms. In the Gaussian case (as opposed to the more general sub-gaussian case, discussed in Theorem 6.7), an *exact* Deterministic Equivalent for the SCM resolvent can be obtained. This can be used to provide a very simple example of non-uniqueness, as is discussed in the following remark.

Remark 6.11 (**Deterministic Equivalents for Gaussian inverse SCM**)**.** A very simple example of Deterministic Equivalents is the following. 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, i.e., $\mathbf{Z}_{ij} \sim \mathcal{N}(0,1)$. In this case, the inverse[18] $\hat{\mathbf{C}}^{-1}$ is known to follow the inverse-Wishart distribution [22] with p degrees of freedom and scale matrix $\mathbf{C}^{-1}$, such that

$$\mathbb{E}[\hat{\mathbf{C}}^{-1}] = \frac{n}{n-p-1}\mathbf{C}^{-1}, \tag{6.33}$$

for $n \geq p+2$. On the other hand, it follows from Theorem 6.5 by taking $z = 0$ in Equation (6.7) that[19]

$$\mathbb{E}[\mathbf{Q}(z)] \leftrightarrow \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p = \frac{n}{n-p}\mathbf{I}_p \tag{6.34}$$

[17] More generally, the basic issue is that, since it corresponds to an inverse, the expectation of the resolvent $\mathbb{E}[\mathbf{Q}(z)]$ is often much *less accessible*, when compared to the expectation $\mathbb{E}[\hat{\mathbf{C}}]$, unless $\hat{\mathbf{C}} \simeq \mathbb{E}[\hat{\mathbf{C}}]$ in a fairly strong sense (which happens in the classical regime).

[18] In the Gaussian setting, the sample covariance $\hat{\mathbf{C}}$ is known to be invertible with probability one if $n \geq p$ and $\mathbf{C}$ is invertible.

[19] Formally, neither Theorem 6.5 nor Theorem 6.7 holds for $z = 0$ and an arbitrary choice of p/n, since we have assumed that $|z| > 0$ in the proof. This is, however, not an issue in the Gaussian setting, in which case the explicit inverse Wishart moments can be used to replace the "rough" control on the $\|\mathbf{Q}(z)\|$ for $z = 0$.

with $m(z) = \frac{1}{1-c} = \frac{n}{n-p}$. Equation (6.34) is an approximation (a "first-order" characterization) of the explicit form in Equation (6.33), for $n, p \gg 1$ and $\mathbf{C} = \mathbf{I}_p$. This example also illustrates that the Deterministic Equivalents are not unique: we could replace the "-1" in denominator of Equation (6.33) by any constant $C' \ll n, p$ to obtain another (equally correct) Deterministic Equivalent.

Bibliography

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley Series in Probability and Statistics. New York: Wiley, 2003.

[2] Zhidong Bai and Jack W. Silverstein. "No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices". In: *The Annals of Probability* 26.1 (1998), pp. 316–345.

[3] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. 2nd ed. Vol. 20. Springer Series in Statistics. Springer-Verlag New York, 2010.

[4] Patrick Billingsley. *Probability and Measure*. 3rd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, 2012.

[5] Charles Bordenave and Djalil Chafaï. "Modern Aspects of Random Matrix Theory". In: *Proceedings of Symposia in Applied Mathematics* (2014), pp. 1–34.

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

[7] Chandler Davis and W. M. Kahan. "Some new bounds on perturbation of subspaces". In: *Bulletin of the American Mathematical Society* 75.4 (1969), pp. 863–868.

[8] M. Derezinski and M. W. Mahoney. "Determinantal Point Processes in Randomized Numerical Linear Algebra". In: *Notices of the AMS* 68.1 (2021), pp. 34–45.

[9] Michal Derezinski et al. "Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update". In: *Advances in Neural Information Processing Systems*. Nov. 2021.

[10] Michal Derezinski et al. "Sparse sketches with small inversion bias". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1467–1510.

[11] Petros Drineas and Michael W. Mahoney. "RandNLA: randomized numerical linear algebra". In: *Communications of the ACM* 59.6 (2016), pp. 80–90.

[12] Carl Eckart and Gale Young. "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1.3 (1936), pp. 211–218.

[13] Massimo Franceschet. "PageRank". In: *Communications of the ACM* 54.6 (2011), pp. 92–101.

[14] Géza Freud. *Orthogonal polynomials*. Elsevier, 2014.

[15] *Generalized Inverses*. CMS Books in Mathematics. New York: Springer-Verlag, 2003.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Ed. by Francis Bach. Adaptive Computation and Machine Learning series. MIT Press, 2016.

[17] 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.

[18] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. 2nd ed. Cambridge University Press, 2012.

[19] Cosme Louart and Romain Couillet. "Concentration of Measure and Large Random Matrices with an application to Sample Covariance Matrices". In: *arXiv* (2018).

[20] 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.

[21] Vladimir A Marcenko and Leonid Andreevich Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Mathematics of the USSR-Sbornik* 1.4 (1967), p. 457.

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

[23] Song Mei and Andrea Montanari. "The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve". In: *Communications on Pure and Applied Mathematics* (2021).

[24] L Mirsky. "Symmetric Gauge Functions And Unitarily Invariant Norms". In: *The Quarterly Journal of Mathematics* 11.1 (1960), pp. 50–59.

[25] Andrew Ng, Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in Neural Information Processing Systems*. Vol. 14. NIPS'02. MIT Press, 2002, pp. 849–856.

[26] Walter Rudin. *Principles of Mathematical Analysis*. 3rd ed. Vol. 3. International Series in Pure and Applied Mathematics. McGraw-Hill Education, 1964.

[27] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.

[28] H. S. Seung, H. Sompolinsky, and N. Tishby. "Statistical mechanics of learning from examples". In: *Physical Review A* 45.8 (1992), pp. 6056–6091.

[29] Jack W. Silverstein and Zhidong Bai. "On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices". In: *Journal of Multivariate Analysis* 54.2 (1995), pp. 175–192.

[30] Elias M. Stein and Rami Shakarchi. *Fourier Analysis: An Introduction*. Princeton University Press, Feb. 2011.

[31] Elias M Stein and Rami Shakarchi. *Functional Analysis, Introduction to Further Topics in Analysis*. 2012.

[32] Gabor Szeg. *Orthogonal polynomials*. Vol. 23. American Mathematical Soc., 1939.

[33] Terence Tao. *Topics in Random Matrix Theory*. Vol. 132. Graduate Studies in Mathematics. 2012.

[34] E. C. Titchmarsh. *The Theory of Functions*. New York, NY, USA: Oxford University Press, 1939.

[35] Aad W. Van der Vaart. *Asymptotic Statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[36] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[37] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed Sensing: Theory and Applications*. Ed. by Yonina C. Eldar and GittaEditors Kutyniok. Cambridge University Press, 2012, 210–268.

[38] Martin J. Wainwright. *High-Dimensional Statistics: : A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[39] Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. "The statistical mechanics of learning a rule". In: *Reviews of Modern Physics* 65.2 (1993), pp. 499–556.

[40] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1-3 (1987), pp. 37–52.

[41] Forrest W. Young. *Multidimensional Scaling: History, Theory, and Applications*. New York: Psychology Press, May 2013.

[42] Y. Yu, T. Wang, and R. J. Samworth. "A useful variant of the Davis–Kahan theorem for statisticians". In: *Biometrika* 102.2 (2015), pp. 315–323.

Appendix A

Technical Results and Lemmas

Here, we describe important technical results that we use.

Borel–Cantelli lemma. Borel–Cantelli lemma is among the most commonly used results in probability, and asserts that if the sum of the probabilities of a sequence of events $\{A_n\}$ is finite, then the set of all outcomes that are "repeated" infinitely many times *must* occur with probability zero.

Theorem A.1 (Borel–Cantelli lemma)**.** *For a sequence of events $A_1, A_2, \ldots$, if $\sum_{n=1}^{\infty} \Pr(A_n) < \infty$, then $\Pr(\limsup_{n\to\infty} A_n) = 0$.*

We will use Borel–Cantelli lemma to prove almost sure convergence of random quantities. For example, if the sequence of random variables $x_1, x_2, \ldots$ satisfy $\Pr(x_n = O(n^{-1/2})) = n^{-2}$, then by the fact that $\sum_{n=1}^{\infty} \Pr(x_n = O(n^{-1/2})) = \pi^2/6 < \infty$ and Borel–Cantelli lemma in Theorem A.1, we have that $x_n \to 0$ almost surely as $n \to \infty$.

Vitali's convergence theorem. Vitali's convergence theorem is generalization of the dominated convergence theorem, which gives a sufficient condition under which limits and integrals of a sequence of functions can be interchanged. Among other things, it gives a sufficient condition for the convergence of expected values of random variables.

Theorem A.2 (Vitali's convergence theorem [34])**.** *Let $f_1, f_2, \ldots$ be a sequence of functions, analytic on a region $D \subset \mathbb{C}$, such that $|f_n(z)| \leq M$ uniformly on n and $z \in D$. Further assume that $f_n(z_j)$ converges for a countable set $z_1, z_2, \ldots \in D$ having a limit point inside D. Then $f_n(z)$ converges uniformly in any region bounded by a contour interior to D. This limit is furthermore an analytic function of z.*

We will heavily exploit Vitali's convergence theorem to study the behavior of resolvents $\mathbf{Q}_{\mathbf{M}}(z)$ and of Stieltjes transforms near the real axis (where it is almost singular but of utmost interest) by instead studying its properties far from the real axis (where it is mathematically more convenient). The theorem is particularly interesting as it states that the knowledge of f_n at a *countable* number of points $z_1, z_2, \ldots$ is sufficient to fully characterize the limit f. In practice, we will use this property when proving convergence of functionals $f_n(z) = g(\mathbf{Q}_{\mathbf{M}}(z) - \bar{\mathbf{Q}}(z)) \to 0$ of *random* resolvents $\mathbf{Q}_{\mathbf{M}}(z)$ to *deterministic* equivalents $\bar{\mathbf{Q}}(z)$ (here n is the growing size of the resolvents). For example, if $f_n(z_j) \to 0$ almost surely for *each* $z_1, z_2, \ldots$, then by the countable union of probability one events, $f_n(z_j) \to 0$ with probability one *uniformly* on the set $\{z_1, z_2, \ldots\}$, and by Vitali we obtain that $f_n(z) \to 0$ with probability one uniformly on a possibly very large subset of $\mathbb{C}$.

Weyl's inequality. Weyl's inequality is a result that can be used to estimate the eigenvalues of a perturbed Hermitian matrix.

Lemma A.3 (Weyl's inequality, [18, Theorem 4.3.1])**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ be symmetric matrices and let the respective eigenvalues of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A}+\mathbf{B}$ be arranged in decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Then, for all $i \in \{1, \ldots, p\}$,*

$$\lambda_{i+j-1}(\mathbf{A}) + \lambda_{p+1-j}(\mathbf{B}) \leq \lambda_i(\mathbf{A}+\mathbf{B}) \leq \lambda_{i-j}(\mathbf{A}) + \lambda_{j+1}(\mathbf{B}) \tag{A.1}$$

In particular,

$$\max_{1\leq i\leq p} |\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{B})| \leq \|\mathbf{A}-\mathbf{B}\|_2.$$

We will use Weyl's inequality to bound the difference between eigenvalues of two matrices, using the spectral norm of their matrix difference.

Davis-Kahan lemma. The Davis-Kahan lemma is a result that uses the eigengap to show how eigenspaces of a matrix change under perturbation. The following is a special case of it.

Lemma A.4 (Davis-Kahan lemma, [7])**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ be symmetric matrices and let the respective eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ be arranged in decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Then,*

$$\sin\theta\left(\mathbf{u}_i(\mathbf{A}), \mathbf{u}_i(\mathbf{B})\right) \leq \frac{\|\mathbf{A}-\mathbf{B}\|_2}{\min\{|\lambda_{i-1}(\mathbf{A}) - \lambda_i(\mathbf{B})|, |\lambda_{i+1}(\mathbf{A}) - \lambda_i(\mathbf{B})|\}} \tag{A.2}$$

for $\sin\theta(\mathbf{u}_1, \mathbf{u}_2) \equiv \sqrt{1-(\mathbf{u}_1^\mathsf{T}\mathbf{u}_2)^2}$, and $\mathbf{u}_i(\mathbf{A}), \mathbf{u}_i(\mathbf{B})$ the eigenvector that corresponds to the eigenvalue of $\lambda_i(\mathbf{A})$ and $\lambda_i(\mathbf{B})$, respectively. The right-hand side bound may depend only on the eigengap *of either $\mathbf{A}$ or $\mathbf{B}$, at the price of a multiplicative factor of two, see [42].*

We will use the Davis-Kahan lemma to bound the angle, as well as the difference in Euclidean norm, between eigenvectors of two matrices, using the spectral norm of their difference.

Woodbury identity and rank-1 perturbation lemma. The Woodbury identity is a result that relates the inverse of a rank-k perturbation of a matrix to a rank-k correction to the inverse of the original matrix. As such, it allows cheap formal computation of inverses and solutions to linear equations.

Lemma A.5 (Woodbury identity)**.** *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,

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

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

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 Lemma A.5 leads to the following rank-1 perturbation lemma for the resolvent of $\mathbf{M}$.

Lemma A.6 (Rank-1 perturbation lemma, [29, Lemma 2.6])**.** *For* $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p\times p}$ *symmetric,* $\mathbf{u} \in \mathbb{R}^p$*,* $\tau \in \mathbb{R}$ *and* $z \in \mathbb{C} \setminus \mathbb{R}$*,*

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

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

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

We will use these results to perform "leave-one-out" type analysis and obtain a self-consistent equation to retrieve Deterministic Equivalents for (random) resolvent matrices.

Resolvent identities. The resolvent identities allow one to manipulate the difference and products involving resolvent matrices (or inverse of matrices).

Lemma A.7 (Resolvent identity)**.** *For invertible matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, we have*

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}.$$

Proof of Lemma A.7. This can be easily checked by multiplying both sides on the left by $\mathbf{A}$ and on the right by $\mathbf{B}$. □

Lemma A.8 (Resolvent trick)**.** *For* $\mathbf{A} \in \mathbb{R}^{p\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{n\times p}$*, we have*

$$\mathbf{A}(\mathbf{B}\mathbf{A} - z\mathbf{I}_n)^{-1} = (\mathbf{A}\mathbf{B} - z\mathbf{I}_p)^{-1}\mathbf{A},$$

for $z \in \mathbb{C}$ *distinct from* 0 *and from the eigenvalues of* $\mathbf{AB}$*.*

Proof of Lemma A.8. Left-multiply both ends of the equality by $\mathbf{AB} - z\mathbf{I}_p$ to obtain $\mathbf{A} = \mathbf{A}$. □

We will use the above resolvent identities to retrieve Deterministic Equivalents for random resolvent matrices.

For $\mathbf{AB}$ and $\mathbf{BA}$ symmetric, Lemma A.8 is a special case of the more general relation

$$\mathbf{A} \cdot f(\mathbf{B}\mathbf{A}) = f(\mathbf{A}\mathbf{B}) \cdot \mathbf{A},$$

with $f(\mathbf{M}) \equiv \mathbf{U} f(\mathbf{\Lambda})\mathbf{U}^\mathsf{T}$ under the eigen-decomposition $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ and f complex analytic. Since f is analytic, $f(\mathbf{BA}) = \sum_{i=0}^\infty c_i(\mathbf{BA})^i$ for some sequence $\{c_i\}_{i=0}^\infty$ and thus $\mathbf{A} \cdot f(\mathbf{BA}) = \sum_{i=0}^\infty c_i(\mathbf{AB})^i \cdot \mathbf{A} = f(\mathbf{AB}) \cdot \mathbf{A}$.

Sylvester's identity. Sylvester's identity connects the determinant and thus eigenvalues of $\mathbf{AB}$ to those of $\mathbf{BA}$, that is, when the (multiplication) order in a matrix product is swapped.

Lemma A.9 (Sylvester's identity, also known as the Weinstein–Aronszajn identity)**.** *For* $\mathbf{A} \in \mathbb{R}^{p\times n}$*,* $\mathbf{B} \in \mathbb{R}^{n\times p}$ *and* $z \in \mathbb{C} \setminus \{0\}$*,*

$$\det(\mathbf{A}\mathbf{B} - z\mathbf{I}_p) = \det(\mathbf{B}\mathbf{A} - z\mathbf{I}_n)\,(-z)^{p-n}.$$

Proof of Lemma A.9. It suffices to develop the block-matrix determinant (recall that $\det \left(\begin{smallmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{smallmatrix}\right) = \det \mathbf{D} \cdot \det(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) = \det \mathbf{A} \cdot \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})$ when $\mathbf{A}, \mathbf{D}$ are invertible)

$$\det \begin{pmatrix} z\mathbf{I}_p & z\mathbf{A} \\ \mathbf{B} & z\mathbf{I}_n \end{pmatrix} = \det(z\mathbf{I}_p) \cdot \det(z\mathbf{I}_n - \mathbf{B}\mathbf{A}) = \det(z\mathbf{I}_n) \cdot \det(z\mathbf{I}_p - \mathbf{A}\mathbf{B}).$$

This concludes the proof of Lemma A.9. □

We will use Sylvester's identity to obtain self-consistent equations and to retrieve Deterministic Equivalents for (random) resolvent matrices.

Block matrix inversion lemma. An immediate consequence of Sylvester's identity is that $\mathbf{AB}$ and $\mathbf{BA}$ have the same *nonzero* eigenvalues (those nonzero values of z for which both left- and right-hand sides vanish). Thus, say for $n \geq p$, $\mathbf{AB} \in \mathbb{R}^{p\times p}$ and $\mathbf{BA} \in \mathbb{R}^{n\times n}$ have the same spectrum, except for the additional $n-p$ zero eigenvalues of $\mathbf{BA}$. This remark implies the next identity.

Lemma A.10 (Block matrix inversion lemma)**.** *For* $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times n}$, $\mathbf{C} \in \mathbb{R}^{n\times p}$ *and* $\mathbf{D} \in \mathbb{R}^{n\times n}$ *with* $\mathbf{D}$ *invertible, we have*

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{S}^{-1} & -\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{S}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix},$$

where $\mathbf{S} \equiv \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ *is the* Schur complement *(for the block* $\mathbf{D}$*) of* $\left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right)$.[20]

We will use the block matrix inversion lemma in handling inverse/resolvent matrices.

Burkholder inequality. Burkholder inequality is a concentration result concerning (the sum of) martingale difference sequences, and is of particular interest when the independence structure exists but is highly complex for the objective of interest.

Lemma A.11 (Burkholder inequality, [3, Lemma 2.13])**.** *Let* $\{X_i\}_{i=1}^{\infty}$ *be a martingale difference for the increasing* σ*-field* $\{\mathcal{F}_i\}$ *and denote* $\mathbb{E}_k$ *the expectation with respect to* $\mathcal{F}_k$*. Then, for* $k \geq 2$*, and some constant* C_k *only dependent on* k*,*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^k\right] \leq C_k \left(\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}_{i-1}[|X_i|^2]\right]^{k/2} + \sum_{i=1}^{n} \mathbb{E}[|X_i|^k]\right).$$

We will use Burkholder inequality specifically to prove concentration result involving resolvent/inverse matrices. Precisely, denote $\mathbb{E}_i[\mathbf{Q}]$ the expectation of the random matrix $\mathbf{Q}$ conditioned on its first (or last) i columns *inside* the inverse, the sequence $\{(\mathbb{E}_i - \mathbb{E}_{i-1})[\mathbf{Q}]\}_{i=1}^{p}$ forms a martingale difference sequence (of matrices); the fluctuation and concentration of such objects (which in a way extend the notion of series of independent random variables) can be controlled with Burkholder inequality.

[20] The Schur complement $\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ is particularly known for its providing the block determinant formula $\det\left(\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right) = \det(\mathbf{D})\det(\mathbf{S})$, already exploited in the proof of Sylvester's identity, Lemma A.9.