1 Introduction

This chapter discusses fundamentally different mental images of large- versus small-dimensional machine learning through examples of sample covariance and kernel matrices, on both synthetic and real data. Random matrix theory is presented as a flexible and powerful tool to assess, understand, and improve classical machine learning methods in this modern large-dimensional setting.

1.1 Motivation: The Pitfalls of Large-Dimensional Statistics

1.1.1 The Big Data Era: When n Is No Longer Much Larger than p

The big data revolution comes along with the challenging needs to parse, mine, and compress a large amount of large-dimensional and possibly heterogeneous data. In many applications, the dimension p of the observations is as large as – if not much larger than – their number n. In array processing and wireless communications, the number of antennas required for fine localization resolution or increased communication throughput may be as large (today in the order of hundreds) as the number of available independent signal observations [Li and Stoica, 2007, Lu et al., 2014]. In genomics, the identification of correlations among hundreds of thousands of genes based on a limited number of independent (and expensive) samples induces an even larger ratio p/n [Arnold et al., 1994]. In statistical finance, portfolio optimization relies on the need to invest on a large number p of assets to reduce volatility but at the same time to estimate the current (rather than past) asset statistics from a relatively small number n of asset return records [Laloux et al., 2000].

As we shall demonstrate in the following section, the fact that in these problems n is not *much larger* than p annihilates most of the results from standard asymptotic statistics that assume n alone is large [Vaart, 2000]. As a rule of thumb, by "much larger" we mean here that n must be at least 100 times larger than p for standard asymptotic statistics to be of practical convenience (see our argument in Section 1.1.2). Many algorithms in statistics, signal processing, and machine learning are precisely derived from this $n \gg p$ assumption that is no longer appropriate today. A major objective of this book is to cast some light on the resulting biases and problems incurred and to provide a systematic random matrix framework to improve these algorithms.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

Possibly more importantly, we will see in this book that (small p) small-dimensional intuitions at the core of many machine learning algorithms (starting with spectral clustering [Ng et al., 2002, Luxburg, 2007]) may strikingly fail when applied in a simultaneously large n,p setting. A compelling example lies in the notion of "distance" between vectors. Most classification methods in machine learning are rooted in the observation that random data vectors arising from a mixture distribution (say Gaussian) gather in "groups" of close-by vectors in the Euclidean norm. When dealing with large-dimensional data, however, concentration phenomena arise that make Euclidean distances useless, if not counterproductive: Vectors from the *same* mixture class may be further away in Euclidean distance than vectors arising from *different* classes. While classification may still be doable, it works in a rather different way from our small-dimensional intuition. The book intends to prepare the reader for the multiple traps caused by this "curse of dimensionality."

1.1.2 Sample Covariance Matrices in the Large n,p Regime

Let us consider the following example that illustrates a first elementary, yet counterintuitive, result: For simultaneously large n,p, the sample covariance matrix $\hat{\mathbf{C}} \in \mathbb{R}^{p\times p}$ based on n samples $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbf{C})$ is an *entry-wise* consistent estimator of the population covariance $\mathbf{C} \in \mathbb{R}^{p\times p}$ (i.e., $\|\hat{\mathbf{C}}-\mathbf{C}\|_\infty \to 0$ as $p,n \to \infty$ for $\|\mathbf{A}\|_\infty \equiv \max_{ij}|\mathbf{A}_{ij}|$) while overall being an extremely poor estimator in a (more practical) operator norm sense (i.e., $\|\hat{\mathbf{C}}-\mathbf{C}\| \not\to 0$, with $\|\cdot\|$ being the operator norm here). Matrix norms are, in particular, *not* equivalent in the large n,p scenario.

Let us detail this claim, in the simplest case where $\mathbf{C} = \mathbf{I}_p$. Consider a dataset $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n] \in \mathbb{R}^{p\times n}$ of n independent and identically distributed (i.i.d.) observations from a p-dimensional standard Gaussian distribution, that is, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I}_p)$ for $i \in \{1,\ldots,n\}$. We wish to estimate the population covariance matrix $\mathbf{C} = \mathbf{I}_p$ from the n available samples. The maximum likelihood estimator in this zero-mean Gaussian setting is the sample covariance matrix $\hat{\mathbf{C}}$ defined by

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}. \tag{1.1}$$

By the strong law of large numbers, for fixed p, $\hat{\mathbf{C}} \to \mathbf{I}_p$ almost surely as $n \to \infty$, so that $\|\hat{\mathbf{C}}-\mathbf{I}_p\| \xrightarrow{\text{a.s.}} 0$ holds for any standard matrix norm and in particular for the operator norm.

One must be more careful when dealing with the case $n,p \to \infty$ with the ratio $p/n \to c \in (0,\infty)$ (or, from a practical standpoint, n is *not much larger* than p). First, note that the entry-wise convergence still holds since, invoking the law of large numbers again,

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

Besides, by a concentration inequality argument, it can even be shown that

$$\max_{1\le i,j\le p}\left|[\hat{\mathbf{C}}-\mathbf{I}_p]_{ij}\right| \xrightarrow{\text{a.s.}} 0,$$

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

which holds as long as p is no larger than a polynomial function of n, and thus:

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_\infty \xrightarrow{\text{a.s.}} 0.$$

Consider now the case $p > n$. Since $\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ is the sum of n rank-one matrices, the rank of $\hat{\mathbf{C}}$ is *at most* equal to n and thus, being a $p \times p$ matrix with $p > n$, the sample covariance matrix $\hat{\mathbf{C}}$ must be a *singular* matrix having at least $p - n > 0$ null eigenvalues. As a consequence,

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\| \not\to 0$$

for $\|\cdot\|$ the matrix operator (or spectral) norm. This last result actually extends to the general case where $p/n \to c \in (0,\infty)$. As such, matrix norms cannot be considered equivalent in the regime where p is not negligible compared to n. This follows from the fact that the coefficients involved in the *equivalence of norm* relation between the infinity and operator norm *depend on p*; here, for instance, we have that for symmetric matrices $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$.

Unfortunately, in practice, the (nonconverging) operator norm is of more practical interest than the (converging) infinity norm.

Remark 1.1 (On the importance of operator norm)**.** *For practical purposes, this "loss" of norm equivalence for large p raises the question of the relevant matrix norm to consider for a given application. For the purpose of the present book, and for most applications in machine learning, the operator (or spectral) norm is the most relevant. First, the operator norm is the matrix norm induced by the Euclidean norm of vectors. Thus, the study of regression vectors or label/score vectors in classification is naturally attached to the spectral study of matrices. Besides, we will often be interested in the* asymptotic equivalence *of families of large-dimensional symmetric matrices. If* $\|\mathbf{A}_p - \mathbf{B}_p\| \to 0$ *for matrix sequences* $\{\mathbf{A}_p\}$ *and* $\{\mathbf{B}_p\}$*, indexed by their dimension p, then according to Weyl's inequality (see, e.g., Lemma 2.10 in Section 2.2.1),*

$$\max_i \left|\lambda_i(\mathbf{A}_p) - \lambda_i(\mathbf{B}_p)\right| \to 0$$

for $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots$*, the eigenvalues of* $\mathbf{A}$ *in a decreasing order. Besides, for* $\mathbf{u}_i(\mathbf{A}_p)$*, an eigenvector of* $\mathbf{A}_p$ *associated with an* isolated eigenvalue $\lambda_i(\mathbf{A}_p)$ *(i.e., such that* $\min\{|\lambda_{i+1}(\mathbf{A}_p) - \lambda_i(\mathbf{A}_p)|, |\lambda_i(\mathbf{A}_p) - \lambda_{i-1}(\mathbf{A}_p)|\} > \varepsilon$ *for some* $\varepsilon > 0$ *uniformly on p),*

$$\left\|\mathbf{u}_i(\mathbf{A}_p) - \mathbf{u}_i(\mathbf{B}_p)\right\| \to 0.$$

These results ensure that, as far as spectral properties are concerned, $\mathbf{A}_p$ *can be studied* equivalently *through* $\mathbf{B}_p$*. We will often use this argument to investigate intractable random matrices* $\mathbf{A}_p$ *by means of a more tractable "proxy"* $\mathbf{B}_p$*.*

The pitfall that consists in assuming that $\hat{\mathbf{C}}$ *is a valid estimator of* $\mathbf{C}$ *since* $\|\hat{\mathbf{C}} - \mathbf{C}\|_\infty \xrightarrow{\text{a.s.}} 0$ *may thus have deleterious practical consequences when n is not* significantly larger *than p.*

Resuming our discussion of norm convergence, it is now natural to ask whether $\hat{\mathbf{C}}$, which badly estimates $\mathbf{C}$, has a controlled asymptotic behavior. There precisely lay the first theoretical interests of random matrix theory. While $\hat{\mathbf{C}}$ itself does not converge in

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

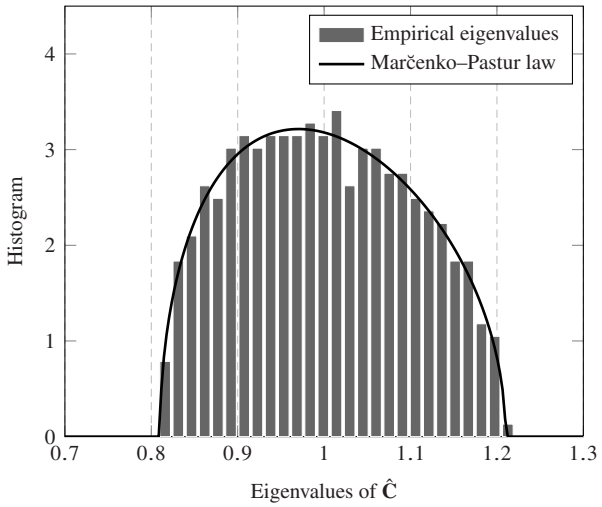


Figure 1.1 Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the Marčenko–Pastur law, for $\mathbf{X}$ having standard Gaussian entries, $p = 500$ and $n = 50000$. Code on web: MATLAB and Python.

any useful way, its eigenvalue distribution does exhibit a traceable limiting behavior [Marčenko and Pastur, 1967, Silverstein and Bai, 1995, Bai and Silverstein, 2010]. The seminal result in this direction, due to Marčenko and Pastur, states that, for $\mathbf{C} = \mathbf{I}_p$, as $n,p \to \infty$, with $p/n \to c \in (0,\infty)$, it holds with probability 1 that the random *discrete eigenvalue/empirical spectral distribution*

$$\mu_p \equiv \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\hat{\mathbf{C}})}$$

converges in law to a nonrandom *smooth* limit, today referred to as the "Marčenko–Pastur law" [Marčenko and Pastur, 1967],

$$\mu(dx) = (1-c^{-1})^+\delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x-E_-)^+(E_+-x)^+}\,dx, \tag{1.2}$$

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

Figure 1.1 compares the empirical spectral distribution of $\hat{\mathbf{C}}$ to the limiting Marčenko–Pastur law given in (1.2), for $p = 500$ and $n = 50000$.

The elementary Marčenko–Pastur result is already quite instructive and insightful.

Remark 1.2 (When is one under the random matrix regime?)**.** *Equation* (1.2) *reveals that the eigenvalues of* $\hat{\mathbf{C}}$*, instead of concentrating at* $x = 1$ *as a large-n alone analysis would suggest, are spread from* $(1-\sqrt{c})^2$ *to* $(1+\sqrt{c})^2$*. As such, the eigenvalues span on a range*

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

This is a slow decaying behavior with respect to $c = \lim p/n$*. In particular, for* $n = 100p$*, in which case, one would expect a sufficiently large number of samples for* $\hat{\mathbf{C}}$ *to properly estimate* $\mathbf{C} = \mathbf{I}_p$*, one has* $4\sqrt{c} = 0.4$*, which is a large spread around*

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

the mean (and true) eigenvalue 1. *This is visually confirmed by Figure 1.1 for* $p = 500$ *and* $n = 50\,000$*, where the histogram of the eigenvalues is nowhere near concentrated at* $x = 1$. *Therefore, random matrix results will be much more accurate than classical asymptotic statistics even when* $n \sim 100p$. *As a telling example, estimating the covariance matrix of each digit from the popular Modified National Institute of Standards and Technology (MNIST) dataset [LeCun et al., 1998], made of no more than* 60 000 *training samples (and thus about* $n = 6\,000$ *samples per digit) of size* $p = 784$*, is likely a hazardous undertaking.*

Remark 1.3 (On universality)**.** *Although introduced here in the context of a Gaussian distribution for* $\mathbf{x}_i$*, the Marčenko–Pastur law applies to much more general cases. Indeed, the result remains valid as long as the* $\mathbf{x}_i$*s have independent normalized entries of zero mean and unit variance (and even beyond this setting, see El Karoui [2009] and Louart and Couillet [2018]). Similar to the law of large numbers in standard asymptotic statistics, this* universality *phenomenon commonly arises in random matrix theory and large-dimensional statistics. We will exploit this phenomenon in the book to justify the wide applicability of the presented results, even to real datasets. See Chapter 8 for more detail.*

1.1.3 Kernel Matrices of Large-Dimensional Data

Another less-known but equally important example of the curse of dimensionality in machine learning involves the loss of relevance of (the notion of) Euclidean distance between large-dimensional data vectors. To be more precise, we will see in the sequel that, *in an asymptotically nontrivial classification setting* (i.e., ensuring that asymptotic classification is neither trivially easy nor impossible), large and numerous data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ extracted from a few-class (say two-class) mixture model tend to be asymptotically at *equal* (Euclidean) distance from one another, irrespective of their corresponding class. Roughly speaking, in this nontrivial setting and under some reasonable statistical assumptions on the $\mathbf{x}_i$s, we have

$$\max_{1 \le i \ne j \le n} \left\{ \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right\} \to 0 \tag{1.3}$$

for some constant $\tau > 0$ as $n, p \to \infty$, *independently of the classes (same or different) of* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ (here the normalization by p is used for compliance with the notations in the remainder of this book and has no particular importance).

This asymptotic behavior is extremely counterintuitive and conveys the idea that classification by standard methods ought *not* to be doable in this large-dimensional regime. Indeed, in the conventional small-dimensional intuition that forged many of the leading machine learning algorithms of everyday use (such as spectral clustering [Ng et al., 2002, Luxburg, 2007]), two data points are assigned to the same class if they are "close" in Euclidean distance. Here we claim that, when p is large, *data pairs are neither close nor far* from each other, regardless of their belonging to the same class or not. Despite this troubling loss of *individual* discriminative power between data pairs, we subsequently show that, thanks to a *collective* behavior of all data

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

belonging to the same (few and thus large) classes, data classification or clustering is still achievable. Better, we shall see that, while many conventional methods devised from small-dimensional intuitions do fail in this large-dimensional regime, some popular approaches, such as the Ng–Jordan–Weiss spectral clustering method [Ng et al., 2002] or the PageRank semisupervised learning approach [Avrachenkov et al., 2012], still function. But the core reasons for their functioning are strikingly different from the reasons of their initial designs, and they often operate far from optimally.

The Nontrivial Classification Regime

To get a clear picture of the source of Equation (1.3), we first need to clarify what we refer to as the "asymptotically nontrivial" classification setting. Consider the simplest scenario of a binary Gaussian mixture classification: Given a training set $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^p$ of n samples independently drawn from the two-class ($\mathcal{C}_1$ and $\mathcal{C}_2$) Gaussian mixture,

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p), \quad \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_p + \mathbf{E}), \tag{1.4}$$

each drawn with probability $1/2$, for some deterministic $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric $\mathbf{E} \in \mathbb{R}^{p \times p}$, both possibly depending on p. In the ideal case where $\boldsymbol{\mu}$ and $\mathbf{E}$ are perfectly known, one can devise a (decision optimal) Neyman–Pearson test. For an unknown $\mathbf{x}$, genuinely belonging to $\mathcal{C}_1$, the Neyman–Pearson test to decide on the class of $\mathbf{x}$ reads

$$(\mathbf{x}+\boldsymbol{\mu})^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-1}(\mathbf{x}+\boldsymbol{\mu}) - (\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu}) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} -\log\det(\mathbf{I}_p+\mathbf{E}). \tag{1.5}$$

Writing $\mathbf{x} = \boldsymbol{\mu} + \mathbf{z}$ for $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, the above test is equivalent to

$$\begin{aligned} T(\mathbf{x}) \equiv\ & 4\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-1}\boldsymbol{\mu} + 4\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-1}\mathbf{z} + \mathbf{z}^\mathsf{T}\left((\mathbf{I}_p+\mathbf{E})^{-1} - \mathbf{I}_p\right)\mathbf{z} \\ & + \log\det(\mathbf{I}_p+\mathbf{E}) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} 0. \end{aligned} \tag{1.6}$$

Since $\mathbf{Uz}$ for $\mathbf{U} \in \mathbb{R}^{p \times p}$, an eigenvector basis of $(\mathbf{I}_p + \mathbf{E})^{-1}$ (and thus of $(\mathbf{I}_p + \mathbf{E})^{-1} - \mathbf{I}_p$), follows the same distribution as $\mathbf{z}$, the random variable $T(\mathbf{x})$ can be written as the sum of p independent random variables. Further assuming that $\|\boldsymbol{\mu}\| = O(1)$ with respect to p, by Lyapunov's central limit theorem (e.g., [Billingsley, 2012, Theorem 27.3]) and the fact that $\mathrm{Var}[\mathbf{z}^\mathsf{T}\mathbf{A}\mathbf{z}] = 2\,\mathrm{tr}(\mathbf{A}^2)$ for symmetric $\mathbf{A} \in \mathbb{R}^{p \times p}$ and Gaussian $\mathbf{z}$, we have, as $p \to \infty$,

$$V_T^{-1/2}(T(\mathbf{x}) - \bar{T}) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\begin{aligned} \bar{T} &\equiv 4\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-1}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{I}_p+\mathbf{E})^{-1} - p + \log\det(\mathbf{I}_p+\mathbf{E}), \\ V_T &\equiv 16\boldsymbol{\mu}^\mathsf{T}(\mathbf{I}_p+\mathbf{E})^{-2}\boldsymbol{\mu} + 2\,\mathrm{tr}\left((\mathbf{I}_p+\mathbf{E})^{-1} - \mathbf{I}_p\right)^2. \end{aligned}$$

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

As a consequence, the classification of $\mathbf{x} \in \mathcal{C}_1$ is asymptotically nontrivial (i.e., the classification error neither goes to 0 nor 1 as $p \to \infty$) if and only if $\bar{T}$ is of the same order as $\sqrt{V_T}$. Considering the (worst-case) scenario where $\mathbf{E} = \mathbf{0}$, we must have $\|\boldsymbol{\mu}\| \geq O(1)$ with respect to p (indeed, if instead $\|\boldsymbol{\mu}\| = o(1)$, the classification of $\mathbf{x}$ is asymptotically impossible).

Under the constraint $\|\boldsymbol{\mu}\| = O(1)$, we move on to consider the case $\mathbf{E} \neq \mathbf{0}$ with the spectral norm constraint $\|\mathbf{E}\| = o(1)$. By a Taylor expansion of both $(\mathbf{I}_p + \mathbf{E})^{-1}$ and $\log\det(\mathbf{I}_p + \mathbf{E})$ around $\mathbf{I}_p$, we obtain

$$\bar{T} = 4\|\boldsymbol{\mu}\|^2 - \frac{1}{2}\operatorname{tr}(\mathbf{E}^2) + o(1);$$
$$V_T = 16\|\boldsymbol{\mu}\|^2 + 2\operatorname{tr}(\mathbf{E}^2) + o(1),$$

which demands $\operatorname{tr}(\mathbf{E}^2)$ to be of order $O(1)$ (same as $\|\boldsymbol{\mu}\|$) so as to have discriminative power. Since $\operatorname{tr}(\mathbf{E}^2) \leq p\|\mathbf{E}\|^2$, with equality if and only if $\mathbf{E}$ is proportional to the identity, that is, $\mathbf{E} = \epsilon\mathbf{I}_p$, one must have $\|\mathbf{E}\| \geq O(p^{-1/2})$. Also, since $O(1) = \operatorname{tr}(\mathbf{E}^2) \leq (\operatorname{tr}\mathbf{E})^2$, we must have $|\operatorname{tr}\mathbf{E}| \geq O(1)$. This allows us to conclude on the following nontrivial classification conditions:

$$\|\boldsymbol{\mu}\| \geq O(1), \quad \|\mathbf{E}\| \geq O(p^{-1/2}), \quad |\operatorname{tr}(\mathbf{E})| \geq O(1), \quad \operatorname{tr}(\mathbf{E}^2) \geq O(1). \tag{1.7}$$

These are the *minimal* conditions for classification in the case of perfectly known means and covariances in the following sense: (i) if none of the inequalities hold (i.e., if the means and covariances from both classes are too close), asymptotic classification must fail and (ii) if at least one of the inequalities is not tight (say if $\|\boldsymbol{\mu}\| \geq O(\sqrt{p})$), asymptotic classification becomes trivial.[1]

We shall subsequently see that (1.7) precisely induces the asymptotic loss of distance discrimination raised in (1.3) but that standard spectral clustering methods based on $n \sim p$ data remain valid.

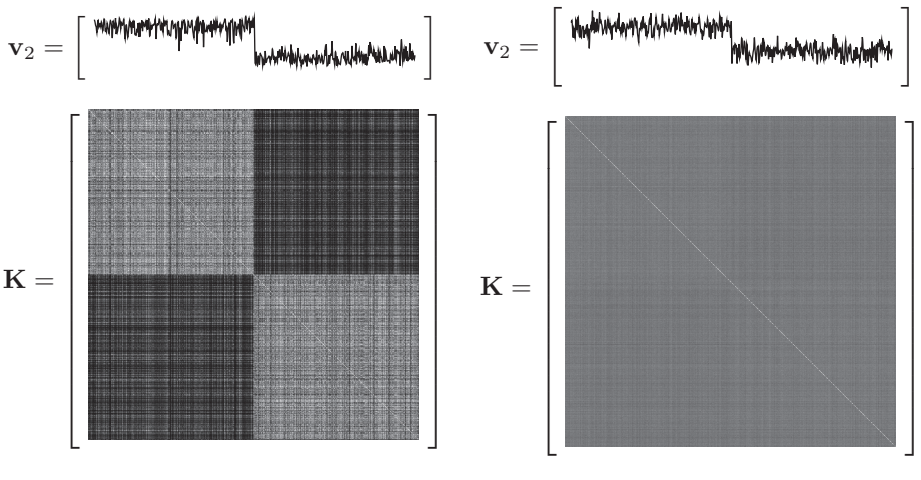
Asymptotic Loss of Pairwise Distance Discrimination

Under the equality case for the conditions in (1.7), consider the (normalized) Euclidean distance between two distinct data vectors $\mathbf{x}_i \in \mathcal{C}_a$ and $\mathbf{x}_j \in \mathcal{C}_b, i \neq j$, given by

$$\frac{1}{p}\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \begin{cases} \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + Ap^{-1}, & \text{for } a = b = 2 \\ \frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2 + Bp^{-1}, & \text{for } a = 1, b = 2, \end{cases} \tag{1.8}$$

[1] It should be noted here that, unlike in computer science, we will stick in this book with the notation $O(\cdot)$ indifferently from the complexity notations $\Omega(\cdot)$, $O(\cdot)$, and $\Theta(\cdot)$. The exact meaning of $O(\cdot)$ will be clear in context. For instance, under computer science notations, Equation (1.7) would be $\|\boldsymbol{\mu}\| \geq \Theta(1)$, $\|\mathbf{E}\| \geq \Theta(p^{-1/2})$, $|\operatorname{tr}(\mathbf{E})| \geq \Theta(1)$, and $\operatorname{tr}(\mathbf{E}^2) \geq \Theta(1)$.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press



(a) $p = 5$ (b) $p = 250$

Figure 1.2 Gaussian kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for **(a)** small- and **(b)** large-dimensional data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, with $\mathbf{x}_1, \ldots, \mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n \in \mathcal{C}_2$ for $n = 5\,000$. Code on web: MATLAB and Python.

where

$$A = \mathbf{z}_i^\mathsf{T} \mathbf{E} \mathbf{z}_i + \mathbf{z}_j^\mathsf{T} \mathbf{E} \mathbf{z}_j - 2\mathbf{z}_i^\mathsf{T} \mathbf{E} \mathbf{z}_j \text{ and}$$

$$B = \mathbf{z}_j^\mathsf{T} (\mathbf{E} + \mathbf{E}^2/4)\mathbf{z}_j - \mathbf{z}_i^\mathsf{T} \mathbf{E} \mathbf{z}_j + 4\|\boldsymbol{\mu}\|^2 + 4\boldsymbol{\mu}^\mathsf{T}(\mathbf{z}_i - \mathbf{z}_j) + o(1)$$

are both of order $O(1)$ $\big($and thus both Ap^{-1} and Bp^{-1} are of order $O(p^{-1})\big)$, while the leading term $\frac{1}{p}\|\mathbf{z}_i - \mathbf{z}_j\|^2$ of (1.8) is of order $O(1)$. As such,

$$\max_{1 \le i \ne j \le n} \left\{ \frac{1}{p} \|\mathbf{z}_i - \mathbf{z}_j\|^2 - 2 \right\} \to 0$$

almost surely as $n, p \to \infty$ (this follows by exploiting the fact that $\|\mathbf{z}_i - \mathbf{z}_j\|^2$ is a chi-square random variable with p degrees of freedom). As a consequence, as previously claimed in (1.3),

$$\max_{1 \le i \ne j \le n} \left\{ \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \tau \right\} \to 0$$

for $\tau = 2$ here. Besides, on a closer inspection of (1.8), we find that, beyond this common value τ of order $O(1)$, the discriminative class information in means $4\|\boldsymbol{\mu}\|^2/p$ and that in covariances $\mathbf{z}_j^\mathsf{T}(\mathbf{E} + \mathbf{E}^2/4)\mathbf{z}_j/p \simeq \operatorname{tr}(\mathbf{E} + \mathbf{E}^2/4)/p$ are both of order $O(p^{-1})$, while by the central limit theorem, $\|\mathbf{z}_i - \mathbf{z}_j\|^2/p = 2 + O(p^{-1/2})$. The class information is thus largely overtaken by the random fluctuations. As a consequence, asymptotically, the pairwise distance $\|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ contains *no* exploitable statistical

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

information (about $\boldsymbol{\mu}$ or $\mathbf{E}$) to distinguish if the $\mathbf{x}_i$ and $\mathbf{x}_j$ vectors belong to the same or different classes.

To visually confirm this joint convergence of the data distances, in Figure 1.2, we display the content of the Gaussian (heat) kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, with $[\mathbf{K}]_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2p)\right)$, and the associated second dominant eigenvector $\mathbf{v}_2$ for a two-class Gaussian mixture $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$, with $\boldsymbol{\mu} = [2;\ \mathbf{0}_{p-1}]$. For a constant $n = 500$, we take $p = 5$ in Figure 1.2(a) and $p = 250$ in Figure 1.2(b).

While the "block-structure" in the case of $p = 5$ of Figure 1.2(a) does agree with the small-dimensional intuition – data vectors from the same class are "closer" to one another in diagonal blocks with larger values (since $\exp(-x/2)$ decreases with x) than in nondiagonal blocks – this intuition collapses when large-dimensional data vectors are considered. Indeed, in the large data setting of Figure 1.2(b), all entries (except obviously on the diagonal) of $\mathbf{K}$ have approximately the same value, which, we now know from (1.3), is $\exp(-1)$.

This is no longer surprising to us. However, what remains surprising in Figure 1.2 at this stage of our analysis is that the eigenvector $\mathbf{v}_2$ of $\mathbf{K}$ seems *not* affected by this (asymptotic) loss of class-wise discrimination of individual distances. And spectral clustering seems to work equally well for $p = 5$ and for $p = 250$, despite the radical and intuitively destructive change in the behavior of $\mathbf{K}$ for $p = 250$.

Explaining Kernel Methods with Random Matrix Theory

The fundamental reason behind this surprising behavior lies in the *accumulated* effect of the $n/2$ small "hidden" informative terms $\|\boldsymbol{\mu}\|^2$, $\operatorname{tr}\mathbf{E}$ and $\operatorname{tr}(\mathbf{E}^2)$ in each class, which collectively "steer" the several top eigenvectors of $\mathbf{K}$. More explicitly, we shall see in the course of this book that the Gaussian kernel matrix $\mathbf{K}$ can be asymptotically expanded as

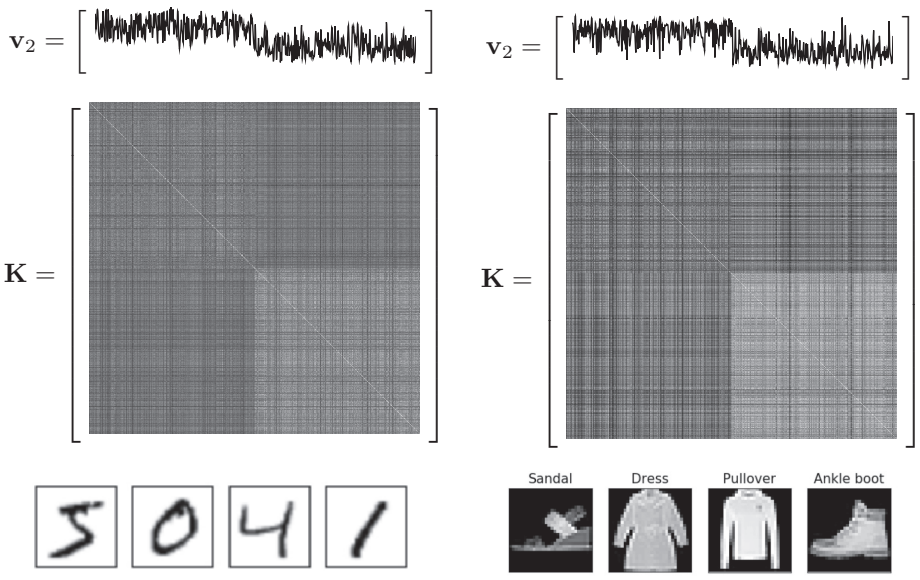
$$\mathbf{K} = \exp(-1)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + f(\boldsymbol{\mu},\mathbf{E})\cdot\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1), \tag{1.9}$$

where $\mathbf{Z} = [\mathbf{z}_1,\ldots,\mathbf{z}_n] \in \mathbb{R}^{p\times n}$ is a Gaussian noise matrix, $f(\boldsymbol{\mu},\mathbf{E}) = O(1)$, and $\mathbf{j} = [\mathbf{1}_{n/2};\ -\mathbf{1}_{n/2}]$ is the class-information "label" vector (as in the setting of Figure 1.2). Here "*" symbolizes extra terms of marginal importance to the present discussion, and $o_{\|\cdot\|}(1)$ represents terms of asymptotically vanishing *operator* norm as $n,p \to \infty$. The important remark to be made here is that

(i) Under this description, $[\mathbf{K}]_{ij} = \exp(-1)(1 + \mathbf{z}_i^\mathsf{T}\mathbf{z}_j/p) \pm f(\boldsymbol{\mu},\mathbf{E})/p + *$, with $f(\boldsymbol{\mu},\mathbf{E})/p \ll \mathbf{z}_i^\mathsf{T}\mathbf{z}_j/p = O(p^{-1/2})$; this is consistent with our previous discussion: The statistical information is *entry-wise* dominated by noise.

(ii) From a *spectral* viewpoint, $\|\mathbf{Z}^\mathsf{T}\mathbf{Z}/p\| = O(1)$, as per the Marčenko–Pastur theorem [Marčenko and Pastur, 1967] discussed in Section 1.1.2 and visually confirmed in Figure 1.1, while $\|f(\boldsymbol{\mu},\mathbf{E})\cdot\mathbf{j}\mathbf{j}^\mathsf{T}/p\| = O(1)$: Thus, *spectrum-wise*, the information stands on even ground with noise.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press



(a) MNIST data (b) Fashion-MNIST data

Figure 1.3 Gaussian kernel matrices $\mathbf{K}$ and the second top eigenvectors $\mathbf{v}_2$ for **(a)** MNIST [LeCun et al., 1998] (class 8 versus 9) and **(b)** Fashion-MNIST [Xiao et al., 2017] data (class 5 versus 7), with $\mathbf{x}_1,\ldots,\mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1},\ldots,\mathbf{x}_n \in \mathcal{C}_2$ for $n = 5\,000$. Code on web: MATLAB and Python.

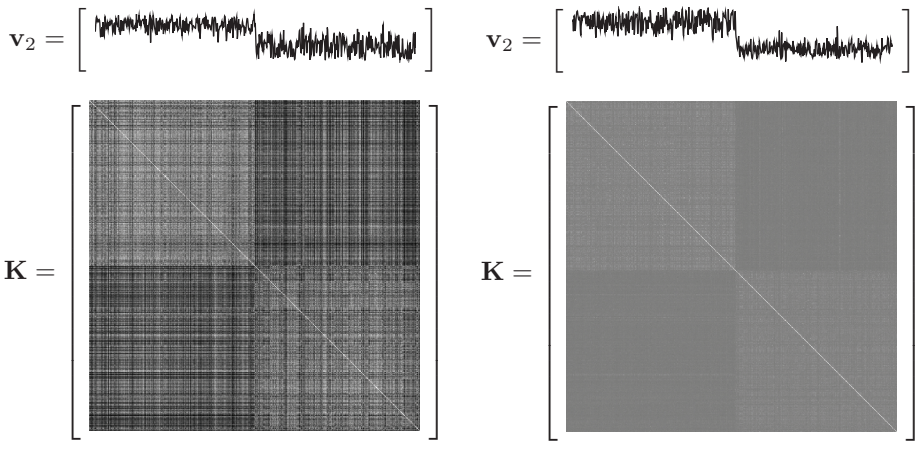
The mathematical magic at play here lies in $f(\boldsymbol{\mu},\mathbf{E}) \cdot \mathbf{j}\mathbf{j}^\mathsf{T}/p$ having entries of order $O(p^{-1})$ while being a low-rank (here unit-rank) matrix: All its "energy" *concentrates* in a single nonzero eigenvalue. As for $\mathbf{Z}^\mathsf{T}\mathbf{Z}/p$, with larger $O(p^{-1/2})$ amplitude entries, it is composed of "essentially independent" zero-mean random variables and tends to be of full rank and *spreads* its energy over its n eigenvalues. Spectrum-wise, both $f(\boldsymbol{\mu},\mathbf{E}) \cdot \mathbf{j}\mathbf{j}^\mathsf{T}/p$ and $\mathbf{Z}^\mathsf{T}\mathbf{Z}/p$ meet on even ground under the nontrivial classification setting of (1.7).

We shall see in Section 4 that things are actually not as clear-cut and, in particular, that not all choices of kernel functions can achieve the same nontrivial classification rates. In particular, the popular Gaussian (radial basis function [RBF]) kernel will be shown to be largely suboptimal in this respect.

Do Real Data Follow Small- or Large-Dimensional Intuitions?

A first glimpse into this riddle, fundamental for the practical design of machine learning algorithms, is provided in Figure 1.3. Similar to Figure 1.2 for synthetic Gaussian data, Figure 1.3 depicts the content of kernel matrices built from the MNIST [LeCun et al., 1998] and Fashion-MNIST data [Xiao et al., 2017], with $p = 28 \times 28 = 784$ and $n = 5\,000$ in both cases. In Figure 1.4, instead of raw data, we display the *features* extracted from popular deep neural networks, such as VGG-16 [Simonyan and Zisserman, 2014] of the more complex CIFAR-10 images (with

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press



(a) VGG-16 features of CIFAR-10 (b) Word2vec features of GoogleNews

Figure 1.4 Gaussian kernel matrices $\mathbf{K}$ and the second dominant eigenvectors $\mathbf{v}_2$ for **(a)** VGG-16 [Simonyan and Zisserman, 2014] features of CIFAR-10 data ("airplane" versus "bird") and **(b)** word2vec [Mikolov et al., 2013] features of GoogleNews-vectors data ("sports" versus "sales"), with $\mathbf{x}_1,\ldots,\mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1},\ldots,\mathbf{x}_n \in \mathcal{C}_2$. Code on web: MATLAB and Python.

$p = 1024$), as well as the so-called "word-embedding" *features* from the popular word2vec method [Mikolov et al., 2013] of the GoogleNews data (with $p = 300$). In all aforementioned cases, we observe a typical large-dimensional behavior (that is similar to Figure 1.2(b) for Gaussian data), not only on raw data but also on efficient features from modern and elaborate machine learning algorithms; even more strikingly, this behavior is *consistently* observed both for image and natural language data, despite their being of a fundamentally different nature. Section 1.2.4, at the end of this introductory chapter, provides first clues that justify why this seemingly unexpected observation (recall again that in the classical motivation behind spectral clustering methods [Ng et al., 2002], we would rather expect a behavior typical of Figure 1.2(a)) on real-world datasets should, in fact, not be a surprise.

1.1.4 Summarizing

In this section, we discussed two simple, yet counterintuitive examples of common pitfalls in learning from large-dimensional data.

In the sample covariance matrix example of Section 1.1.2, we made the important remark of the *loss of equivalence* between matrix norms in the *random matrix regime* where the data (or features) dimension p and their number n are both large and comparable, which is at the source of many seemingly striking empirical observations in modern machine learning. We, in particular, insist that for matrices $\mathbf{A}_n, \mathbf{B}_n \in \mathbb{R}^{n \times n}$ of large sizes,

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

$$\forall i,j,\ [\mathbf{A}_n - \mathbf{B}_n]_{ij} \to 0 \not\Rightarrow \|\mathbf{A}_n - \mathbf{B}_n\| \to 0 \tag{1.10}$$

in the operator norm.

We also realized, from a basic reading of the Marčenko–Pastur theorem, that the random matrix regime arises more often than one may think: While $n/p \sim 100$ may seem a large enough ratio for classical asymptotic statistics to be accurate, random matrix theory is, in general, a far more appropriate tool (with as much as 20% gain in precision for the estimation of the eigenvalues of sample covariances).

In Section 1.1.3, we provided a concrete machine learning application example of the message in (1.10). We saw that, in the practically most relevant scenario of nontrivial (not too easy, not too hard) large data classification, the Euclidean distance between any two data vectors "concentrates" around a constant as in (1.3), regardless of their respective classes. Yet, since again entry-wise convergence $[\mathbf{A}_n]_{ij} \to \tau$ does not imply operator norm convergence $\|\mathbf{A}_n - \tau \mathbf{1}_n \mathbf{1}_n^\mathsf{T}\| \to 0$, we understood that, thanks to a collective effect of the small but similarly "oriented" fluctuations in all the entries, spectral clustering remains valid for large-dimensional problems.

Possibly most importantly, we discovered that the "curse of dimensionality" induced by the counterintuitive behavior of large-dimensional vectors turns into an asset for mathematical analysis. In the sample covariance matrix example, we observed that a random-matrix version of the laws of large numbers arises in the convergence of the eigenvalue distributions of large sample covariance matrices to a deterministic limiting measure. As a matter of fact, as we shall see throughout the book, the very fact that both p and n are large ensures a generally *fast convergence* of most (random) quantities of practical interest for machine learning: By exploiting $np = O(n^2)$, rather than n degrees of freedom, central limit theorems may converge at $O(1/n)$ rate (instead of the classical $O(1/\sqrt{n})$ rate).

This fast convergence rate further induces another important phenomenon, referred to as the *universality*, which ensures the robustness of the random matrix asymptotics to a vast range of distributions. Essentially, as we shall see in more detail later in this book, first- and second-order statistics are often *sufficient* to describe most asymptotic behaviors, even of complicated data models and methods. This is a first (yet not the most convincing) justification of the repeatedly observed – but quite unexpected – good match between random matrix predictions and experiments on real datasets.

In a nutshell, the fundamentally counterintuitive, yet mathematically addressable changes in the behavior of large-dimensional data when compared with small-dimensional data have two major consequences to statistics and machine learning: (i) most algorithms, originally developed under a small-dimensional intuition, are likely to fail (as we shall discover in this book, many of them do) or at least to perform inefficiently and (ii) by benefiting from the extra degrees of freedom offered by large data (in the dimension p), random matrix theory is apt to analyze and improve these methods, but most importantly, it generates a whole new paradigm for large-dimensional learning.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

1.2 Random Matrix Theory as an Answer

1.2.1 Which Theory and Why?

A Point of History

Random matrix theory originates from the work of John Wishart [Wishart, 1928] on the study of the eigenvalues of the matrix $\mathbf{X}\mathbf{X}^\mathsf{T}$ (now referred to as a Wishart matrix) for $\mathbf{X} \in \mathbb{R}^{p\times n}$ with standard Gaussian entries $[\mathbf{X}]_{ij} \sim \mathcal{N}(0,1)$. Wishart managed to determine a closed-form expression for the joint eigenvalue distribution of $\mathbf{X}\mathbf{X}^\mathsf{T}$ for every pair of p,n. Few progress however followed, as matrices with non-Gaussian entries are hardly amenable to similar analysis and, even if they were, the actual study of more elaborate functionals of $\mathbf{X}\mathbf{X}^\mathsf{T}$ is at best cumbersome and often simply intractable.

The works of the physicist Eugene Wigner [Wigner, 1955] gave a new impulse to the theory. Interested in the eigenvalues of symmetric matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ with independent Bernoulli entries (particle spins in his application context), Wigner opted for an *asymptotic* analysis of the eigenvalue distribution, thereby initiating the important and much richer branch of *large-dimensional random matrix theory*. Despite this important inspiration, Wigner exploited standard asymptotic statistics tools (the method of moments) to prove that the *discrete* distribution of the eigenvalues of $\mathbf{X}$ has a *continuous* semicircle looking density in the $n \to \infty$ limit (the now popular semicircular law). This approach was particularly convenient as the limiting law is simple and could be visually anticipated (which is not the case of the next-to-come Marčenko–Pastur limiting distribution of Wishart matrices).

Only until 1967 with the tour-de-force of Marčenko and Pastur [1967] did random matrix theory take a new dimension. Marčenko and Pastur determined the limiting spectral distribution of the sample covariance matrix model $\mathbf{X}\mathbf{X}^\mathsf{T}$ of Wishart but under relaxed conditions: $[\mathbf{X}]_{ij}$ are independent entries with zero mean and unit variance, and additional moment assumptions (all discarded in subsequent works). The independence (or weak dependence) property is key to their proof, which exploits the powerful Stieltjes transform $\frac{1}{p}\operatorname{tr}(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1} = \int(\lambda - z)^{-1}\mu_p(dt)$ of the *empirical spectral distribution* $\mu_p \equiv \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})}$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$, a tool borrowed from operator theory in Hilbert spaces [Akhiezer and Glazman, 2013], rather than the moments $\frac{1}{p}\operatorname{tr}(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T})^k$ (which may not converge since $\mathbb{E}[\mathbf{X}_{ij}^\ell]$ needs not be finite for $\ell > 2$).

The technical approach devised by Marčenko and Pastur was then largely embraced at the turn of the twenty-first century by Bai and Silverstein who, in a series of significant breakthroughs (the most noticeable of which are [Silverstein and Bai, 1995, Bai and Silverstein, 1998]), extended the results in [Marčenko and Pastur, 1967] to an exhaustive study of sample covariance matrices.

In parallel, another approach to limiting spectral analysis of large random matrices emerged as an application example of the *free probability theory* developed by Voiculescu et al. [1992]. Free probability was born as a theory to study random variables in noncommutative algebras, such as the algebra of matrices. Rather than relying on independence assumptions as for the aforementioned Stieltjes transform method,

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

free probability theory relies on a notion of *asymptotic freeness*. In essence, random matrices are asymptotically free if their eigenvector distributions are sufficiently "isotropic" with respect to each other; for instance, independent Gaussian matrices (matrices with independent Gaussian entries) are free, and independent unitary matrices with isotropic eigenvector distributions are free, and a deterministic matrix is free with respect to a Gaussian matrix [Mingo and Speicher, 2017].

Both free probability and the Stieltjes transform approaches have long lived hand-in-hand, and are essentially capable of proving similar results under various assumptions. A classical example, of great importance to this book, is that of *spiked models* (i.e., finite-rank deformations of random matrices, such as the nonzero mean sample covariance $(\mathbf{X} + \boldsymbol{\mu}\mathbf{1}_n^\mathsf{T})(\mathbf{X} + \boldsymbol{\mu}\mathbf{1}_n^\mathsf{T})^\mathsf{T}$ or the rank-one perturbed identity covariance $(\mathbf{I}_p + \ell\mathbf{u}\mathbf{u}^\mathsf{T})^{\frac{1}{2}}\mathbf{X}\mathbf{X}^\mathsf{T}(\mathbf{I}_p + \ell\mathbf{u}\mathbf{u}^\mathsf{T})^{\frac{1}{2}}$ for $\mathbf{X}$ with i.i.d. zero-mean entries) made popular by two key articles [Baik and Silverstein, 2006] and [Benaych-Georges and Nadakuditi, 2012], respectively based on a Stieltjes transform and a free probability approach.

These tools are largely sufficient to cover most of the basic statistical problems in random matrix theory. In particular, the often-called *global regime* of random matrices: Their limiting eigenvalue distribution, the behavior of linear statistics of their eigenvalues or eigenvectors, the position of the outlying eigenvalues in spiked models, etc., are all accessible by either method. However, this is often not the case of the *local regime*: The limiting distribution of a specific eigenvalue (notably the largest and smallest, of practical interest) for which more efforts are, in general, needed. There, researchers have rather resorted to a finite-dimensional analysis of the joint eigenvalue distribution for the Gaussian case (in the spirit of Wishart), and carefully taken the limits of the distribution, exploiting powerful tools such as orthogonal polynomial theory [Johnstone, 2001]. We will not further discuss these approaches in the book, which are rather specific and not of direct use to our applications.

Resolvents, Gaussian Tools, and Concentration of Measure Theory

As we shall see throughout this book, realistic data and feature models necessarily contain rich statistical structures and information patterns (to be extracted by machine learning algorithms). Typical examples include local structures (captured by convolutional filters) in image data, as well as short- and long-term dependences in time series or natural language data. In random matrix terms, this involves dealing with very structured and heterogeneous random matrix models. Although it ebbed and flowed in the past decade, the free probability approach, in general, requires increased effort and advanced techniques to prove the key asymptotic freeness, if possible at all. For this reason (and also because most research and results are available in the Stieltjes transform-related literature), our focus in this book will be on the range of methods surrounding the Stieltjes transform approach.

More exactly, the central object of study in this book is the so-called *resolvent* of the (almost always symmetric, or Hermitian in the complex case) random matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ under investigation, that we shall often denote $\mathbf{Q}_\mathbf{X}(z)$ or simply $\mathbf{Q}(z)$, and

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

that is defined, for all $z \in \mathbb{C}$ not in the eigenspectrum of $\mathbf{X}$ (i.e., not coinciding with an eigenvalue of $\mathbf{X}$), by

$$\mathbf{Q}_{\mathbf{X}}(z) \equiv (\mathbf{X} - z\mathbf{I}_n)^{-1}. \tag{1.11}$$

The resolvent is a rich mathematical object that gives access to:

- *the eigenvalue distribution* $\mu_{\mathbf{X}} \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{X})}$ of $\mathbf{X}$ through the (inverse) Stieltjes transform relation (for all $a,b \notin \{\lambda_1(\mathbf{X}),\ldots,\lambda_n(\mathbf{X})\}$)

$$\int_a^b \mu_{\mathbf{X}}(d\lambda) = \lim_{\epsilon \downarrow 0} \int_a^b \frac{1}{\pi} \Im[m_{\mathbf{X}}(x + \imath\epsilon)]\, dx,$$

  with $\imath$ the imaginary unit and

$$m_{\mathbf{X}}(z) \equiv \int \frac{\mu_{\mathbf{X}}(d\lambda)}{\lambda - z} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda_i(\mathbf{X}) - z} = \frac{1}{n} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z);$$

- *functionals of these eigenvalues* $\frac{1}{n}\sum_{i=1}^{n} f(\lambda_i(\mathbf{X}))$ through Cauchy's integral identity (Theorem 2.2)

$$\frac{1}{n}\sum_{i=1}^{n} f(\lambda_i(\mathbf{X})) = -\frac{1}{2\pi\imath n} \oint_{\Gamma} f(z) \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)\, dz,$$

  for $\Gamma \subset \mathbb{C}$, a positively oriented contour in the complex plane surrounding all the $\lambda_i(\mathbf{X})$s and $f(z)$ complex analytic in a neighborhood of the "inside" of Γ;
- *the eigenvectors and subspaces* of $\mathbf{X}$, again, through Cauchy's integral relation

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

  for $(\lambda_i(\mathbf{X}),\mathbf{u}_i(\mathbf{X}))$, an eigenpair of $\mathbf{X}$ and $\Gamma_{\lambda_i(\mathbf{X})}$, a positively oriented contour surrounding only $\lambda_i(\mathbf{X})$.

As such, the resolvent plays a key role in the analysis of spectral methods, such as (kernel) spectral clustering or graph-based community detection, in which case, the top eigenvectors of some underlying random matrix are exploited.

In addition, the resolvent is a fundamental object that frequently appears in the solutions to linear regression problems (for machine learning applications, in least squares support vector machines, random features and kernel ridge regressions, neural networks, etc.), or to random walk and graph-based semi-supervised learning methods. They will also be shown to appear naturally in not immediately related machine learning problems, such as in large-dimensional nonlinear regression (such as logistic or robust M-regression).

The core of the random matrix approach devised in this book consists in determining, for various statistical models of random matrices $\mathbf{X}$, a *deterministic equivalent* $\bar{\mathbf{Q}}(z)$ for $\mathbf{Q}(z) = \mathbf{Q}_{\mathbf{X}}(z)$, that it is a deterministic matrix $\bar{\mathbf{Q}}(z)$ such that

$$u(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z)) \xrightarrow{\text{a.s.}} 0, \quad \text{or} \quad u(\mathbb{E}[\mathbf{Q}(z)] - \bar{\mathbf{Q}}(z)) \to 0$$

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

for all 1-Lipschitz linear mapping $u\colon \mathbb{R}^{n\times n} \to \mathbb{R}$. Of particular interest are the functions $u(\mathbf{X}) = \frac{1}{n}\operatorname{tr}(\mathbf{AX})$ for $\|\mathbf{A}\| \leq 1$, and $u(\mathbf{X}) = \mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{b}$ for $\|\mathbf{a}\|, \|\mathbf{b}\| \leq 1$.[2]

As an example, in the setting of the Marčenko–Pastur law, where the random matrix of interest is $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ with $\mathbf{X} \in \mathbb{R}^{p\times n}$ having i.i.d. zero mean and unit variance entries, the resolvent

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

admits

$$\bar{\mathbf{Q}}(z) = m_\mu(z)\mathbf{I}_p, \quad m_\mu(z) = \int \frac{\mu(d\lambda)}{\lambda - z}, \quad \text{for } \mu \text{ defined in (1.2)},$$

as a deterministic equivalent. Thus, in particular, $\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) - m_\mu(z) \xrightarrow{\text{a.s.}} 0$ and $\mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{b} - m_\mu(z)\mathbf{a}^\mathsf{T}\mathbf{b} \xrightarrow{\text{a.s.}} 0$ for deterministic $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$ of bounded Euclidean norm.

Consequently, the resolvent (and Stieltjes transform) approach simultaneously involves notions from three distinct mathematical areas:

- *linear algebra*, and particularly the exploitation of inverse matrix lemmas, the Schur complement, interlacing, and low-rank perturbation identities [Horn and Johnson, 2012];
- *complex analysis* (the resolvent $\mathbf{Q}(z)$ is a complex analytic matrix-valued function), and particularly the theory of analytic functions, contour integrals, and residue calculus [Stein and Shakarchi, 2003];
- *probability theory*, and, most specifically, notions of convergence, central limit theory, and the method of moments [Billingsley, 2012]. Depending on the underlying random matrix assumptions (independence of entries, Gaussianity, concentration properties), different random matrix-adapted techniques (among others and variations) will be discussed in this book: the Gaussian tools developed by Pastur, relying on Stein's lemma and the Nash–Poincaré inequality [Pastur and Shcherbina, 2011], the Bai–Silverstein inductive method [Bai and Silverstein, 2010], the concentration of measure framework developed by Ledoux [2005] and applied to random matrix endeavors successively by El Karoui [2009], Vershynin [2012], and Louart and Couillet [2018], or the double leave-one-out approach devised by El Karoui et al. [2013].

The aforementioned tools are, in general, used together with a *perturbation approach* in the sense that they exploit the fact that, by eliminating a row or a column (say, here both row and column i) of a large random matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ to obtain $\mathbf{X}_{-i} \in \mathbb{R}^{(n-1)\times(n-1)}$, the resulting resolvent $\mathbf{Q}_{-i}(z) = (\mathbf{X}_{-i} - z\mathbf{I}_{n-1})^{-1}$ can be related to the original resolvent $\mathbf{Q}(z)$ through both linear algebraic relations and *asymptotically*

[2] Here, $\mathbf{A}$ and $\mathbf{a}, \mathbf{b}$ must be understood as "sequences" of deterministic matrices (or vectors) of growing size but with controlled norm; in particular, $\mathbf{A}$ and $\mathbf{a}, \mathbf{b}$, being deterministic, *cannot* depend on $\mathbf{X}$ (in which case, the convergence results may fail: take for instance $\mathbf{a} = \mathbf{b}$ some eigenvector of $\mathbf{X}$ to be convinced).

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

comparable statistical behaviors. For instance, in the case of symmetric $\mathbf{X}$ with i.i.d. (and properly normalized) entries, it is not difficult to show that $m_{\mathbf{X}}(z) = m_{\mathbf{X}_{-i}}(z) + O(n^{-1})$.

In this regard, Pastur's Gaussian method manages, for models of $\mathbf{X}$ involving Gaussianity (e.g., $\mathbf{X}$ has Gaussian entries or its entries are functions of Gaussian random variables), to obtain asymptotic relations for $\mathbb{E}\mathbf{Q}(z)$. Interpolation methods may then be used to extrapolate the results beyond the Gaussian setting. The Bai–Silverstein inductive method, on the contrary, is not restricted to matrices with Gaussian entries but is restricted to the specific analysis of either trace forms $\mathrm{tr}\,\mathbf{A}\mathbf{Q}(z)$ or bilinear forms $\mathbf{a}^{\mathsf{T}}\mathbf{Q}(z)\mathbf{b}$ that need be treated individually (it also suffers to handle exotic forms of dependence within $\mathbf{X}$). The concentration of measure approach is quite versatile: by merely restricting the matrix under study to be constituted of *concentrated random vectors* (so, in particular, Lipschitz maps of standard Gaussian random vectors or of vectors with i.i.d. entries), it allows one to study simultaneously the fluctuations of all linear functionals of $\mathbf{Q}(z)$ under light conditions on $\mathbf{X}$.

1.2.2 The Double Asymptotics: Turning the Curse of Dimensionality into a Dimensionality Blessing

Why Random Matrix Theory to Study the Large *n,p* Regime?

Although we have previously made a point that modern data processing and learning involve large dimensions (numerous data, large sample sizes, large number of system parameters), and that large-dimensional statistics are a natural class of mathematical tools to turn to, why should one invest in random matrix theory rather than, say, statistical physics,[3] nonasymptotic random matrix theory,[4] or compressive sensing?[5] Large-dimensional random matrix theory, as we introduce it in this book, has two key

[3] Statistical physics and statistical mechanics are powerful tools to map large-dimensional data problems into physics-inspired problems of "interacting particles" [Mézard and Montanari, 2009]. In the early 2000s, statistical physics has brought inspiring ideas and powerful (but unfortunately often unreliable, since nonrigorous) tools for the analysis of wireless communication and information-theoretic problems, before being caught up by added solid and versatile mathematical techniques. Today, statistical physics has an edge on the study of *sparse* (graph-based) machine learning problems for which random matrix theory still struggles to offer a sound theory.

[4] The recent field of nonasymptotic random matrix theory is based on concentration inequality approach and aims, as such, to provide bounds rather than exact (deterministic) asymptotics on various random matrix quantities [Vershynin, 2018]. This set of concentration inequalities should not to be confused with the *concentration of measure theory* [Ledoux, 2005]: Concentration inequalities form a restricted subset of the theory by proving statistical bounds on specific quantities.

[5] Compressive sensing revolves around the assumption that large (p)-dimensional data often arise from a manifold in $\mathbb{R}^p$ of much lower intrinsic dimension: Under this assumption, the curse of dimensionality (when $p \sim n$ or even $p \gg n$) vanishes if one manages to retrieve the (often unknown) low-dimensional manifold. As an aftermath of the seminal work by Candes and Tao [2005], compressive sensing was possibly the first major breakthrough in the modern field of large-dimensional statistical machine learning.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

distinctive features, making it simultaneously more powerful and versatile than these alternative tools:

(i) Unlike nonasymptotic random matrix theory and compressive sensing methods, which mostly aim at *bounding* key quantities (from a rather *qualitative* standpoint), large-dimensional random matrix theory is able to provide *precise and quantitative* (asymptotically exact) approximations for a host of quantities, defined as functionals of random matrices. As a matter of fact, nonasymptotic random matrix theory is more flexible in its not constraining the system dimensions (p, n) and latent variables (data statistics, model hyperparameters) to increase at a controlled rate. Large-dimensional random matrix theory, on the contrary, imposes a controlled growth on the dimensions, *and consequently*, on the model statistics to enforce nontrivial limiting behavior. The ensuing drawback of this allowed flexibility is that only qualitative bounds can be obtained on the system behavior, which at best provides "rules of thumbs" and order of magnitudes on the performance of given algorithms. Large-dimensional random matrix theory, by providing *exact* asymptotics, allows one to finely track the system behavior and opens the possibility to improve its (also fully traced) performance.

(ii) Modern advances in large-dimensional random matrix theory, as opposed to statistical physics notably, further provide results for rather generic and complex system models: matrix models involving nonlinearities (kernels, activation functions), structural data dependence (nonidentity covariances, heterogeneous mixture models, models of concentrated random vectors with strong nonlinear dependence). These key features bring the random matrix tools much closer to practical settings and algorithms. As such, not only does random matrix theory provide a precise understanding of the behavior of key algorithms in machine learning, but it also predicts their behavior when applied to realistic data models.

These two advantages are decisive to the analysis, improvement, and proposition of new machine learning algorithms.

The Case of Machine Learning

The major technical difficulty that has long held many machine learning away from *precise quantitative* analysis and theoretical comprehension relates to the *nonlinearity* involved in feature extraction (nonlinear kernels, nonlinear activation functions in neural networks), to the *implicit* nature of some methods (as simple as the logistic regression), and eventually to the difficulty of a proper (statistical) modeling of *complex* realistic data of various natures (starting with natural images).

An all-encompassing example of these difficulties could be summarized in the following classical problem:

Problem. Determine the *exact* classification performance of logistic regression for n independent observations of p-dimensional (random) feature vectors extracted from a set of two-class images (say, images of dogs versus images of cats).

In the conventional wisdom of statistical machine learning, one *cannot* conceive to solve this problem in an *exact* and *qualitative* manner: the input data (real images)

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

are not easily modeled, the nonlinear features extracted from those data are complex mathematical objects (even in the case where the original data could be modeled as multivariate Gaussian random vectors), and the logistic regression is an implicit optimization method not easily amenable to explicit mathematical analysis.

We shall demonstrate throughout this book that random matrix theory provides a satisfying answer to all these difficulties at one fell swoop and can actually *solve* the **Problem**. This is made possible by the powerful joint *universality* and *determinism* effects brought by large-dimensional data models and treatments.

Specifically, in the random matrix regime where n,p grow large at a controlled rate, the following key properties arise:

- *fast asymptotic determinism*: the law of large numbers and the central limit theorem tell us that the average of n i.i.d. random variables converges to a deterministic limit (e.g., the expectation) at an $O(1/\sqrt{n})$ speed. By gathering independence (or degrees of freedom) both in the sample dimension p and size n, functionals of large random matrices (even mathematically involved functionals, such as the average of functions of their eigenvalues) also converge to deterministic limits, but at an increased speed of up to $O(1/\sqrt{np})$ which, for $n \sim p$, is $O(1/n)$. In machine learning problems, performance may be expressed in terms of misclassification rates or regression errors (i.e., averaged statistics of sometimes involved random matrix functionals) and can thereby be predicted with high accuracy, even for not too large datasets;
- *universality with respect to data models*: similarly, again, consistently with the law of large numbers and the central limit theorem in the large-n alone setting, the above asymptotic deterministic behavior at large n,p is, in general, *independent* of the underlying distribution of the random matrix entries. This phenomenon, referred to in the random matrix literature as *universality*, predicts notably that the asymptotic statistics of even complex machine learning procedures depend on the input data only via the first- and second-order statistics; this is a major distinctive feature when compared to the fixed-p and large-n regime, where the asymptotic performance of algorithms, when accessible, would, in general, depend on the exact p-dimensional distribution of the data;[6]
- *universality with respect to algorithm nonlinearities*: when nonlinear methods are considered, the nonlinear function f (e.g., the kernel function or the activation function) gets involved in the large-dimensional machine learning algorithm performance only via a few parameters (e.g., its derivatives $f(\tau), f'(\tau), \ldots$ at a precise location τ, its "moments" $\int f^k \mu$ with respect to the Gaussian measure μ, or more elaborate scalars solution to a fixed-point equation involving f). For instance, in the case of kernel random matrices of the type $f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)$, only

[6] Compare, for instance, Luxburg et al. [2008] on the fixed-p and large-n asymptotics of spectral clustering (the main result of which contains nonlinear expressions of the input data distribution) to Couillet and Benaych-Georges [2016] on the large p, n asymptotics of the same problem (the main result of which only involves linear and quadratic forms of the statistical mean and covariances of the data, irrespective of the input data distribution, as further confirmed by Seddik et al. [2019]).

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

the first three successive derivatives of the kernel function f at the "concentration" point $\tau = \lim_p \|\mathbf{x}_i - \mathbf{x}_j\|^2/p$ matter; the performance of random neural networks depends on the nonlinear activation function $\sigma(\cdot)$ solely through its first Hermite coefficients (i.e., its Gaussian moments); in implicit optimization schemes (such as logistic regression), the solution "concentrates" with predictable asymptotics, which, despite the nonlinear and implicit nature of the problem, only depend on a few scalar parameters of the logistic loss function. This, together with the asymptotic deterministic behavior of the linear (eigenvalue or eigenvector) statistics discussed above, gives access to the performance of a host of *nonlinear* machine learning algorithms.

- *tractable real data modeling*: possibly, the most important aspect of large-dimensional random matrix analysis in machine learning practice relates to the counterintuitive fact that, as p,n grow large, machine learning algorithms tend to treat real data *as if they were mere Gaussian mixture models*. This statement, to be discussed thoroughly in the subsequent sections, is both supported by empirical observations (with most theoretical findings derived for Gaussian mixtures observed to fit the performances retrieved on real data) and by the theoretical fact that some extremely realistic datasets (in particular, artificial images created by the popular generative adversarial networks, or GANs) are by definition *concentrated random vectors*, which are: (i) amenable to (and, in fact, extremely well-suited for) random matrix analysis, and (ii) proven to behave as if they were mere Gaussian mixtures.

In a word, in large-dimensional problems, data no longer "gather" in groups and do not really "spread" all over their large ambient space neither. But, by accumulation of degrees of freedom, they rather concentrate within a thin lower-dimensional "layer." Each scalar observation of the data, even through complicated functions (regressors, classifiers for machine learning applications), tends to become deterministic, predictable, and simple functions of first-order statistics of the data distribution. Random matrix theory exploits these effects and is thus able to answer seemingly inaccessible machine learning questions.

1.2.3 Analyze, Understand, and Improve Large-Dimensional Machine Learning Methods

One of the first elementary objectives of this book is to demonstrate that, in a large-dimensional and numerous data setting, many standard low-dimensional machine learning intuitions tend to collapse. As a result, many of the algorithms originally designed for small-dimensional data fail to perform as expected. Some of these algorithms will be shown to remain valid, but for rather unexpected reasons. And some of them will be proven suboptimal, quite largely so sometimes. Finally, some of them will be shown to completely fail to meet their objectives and in need of an adaptation or a complete change of paradigm.

In a second part, the book will further show that this "large-dimensional" regime, which one may think synonymous to thousands or millions in dimension and sample size, is in reality already visible in much smaller data sizes than the earliest researchers

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

in applied random matrix theory could anticipate. And, more importantly, that a large class of "real data" naturally falls under the random matrix theory umbrella.

Our argumentation line and every single treatment of machine learning algorithm analysis and improvement proceed along the following steps: One needs to (i) conceive the limitations of low-dimensional intuitions and understand the reach of the very different large-dimensional intuitions, (ii) capture the behavior of the main mathematical objects at play in machine learning method on large-dimensional models so as to (iii) include these objects in a mathematical framework for performance analysis, and (iv) foresee means of improvement based on the newly acquired large-dimensional intuitions and mathematical understanding.

In the remainder of this subsection, we will illustrate the above four-step methodology with the examples of kernel methods and the very related random feature maps (which may alternatively be seen as a two-layer neural network model with random first-layer weights).

From Low- to Large-Dimensional Intuitions

Most of the manuscript focuses on large-dimensional data vectors or graph models. By large-dimensional, we refer to random vectors $\mathbf{x} \in \mathbb{R}^p$ "built from" numerous (of order $O(p)$) degrees of freedom. That is, as opposed to the compressive sensing paradigm [Donoho, 2006], we do not impose the existence of a low-dimensional representation of the data.[7]

From this viewpoint, the simplest mixture data model is the symmetric binary Gaussian mixture model $\mathbf{x} \sim \mathcal{N}(\pm\boldsymbol{\mu}, \mathbf{I}_p)$. As we saw previously, for p small (say, $p = 2$ or $p = 3$), classifying n samples of the mixture is easily visualized as grouping two stacks of data: one gathered around $\boldsymbol{\mu} \in \mathbb{R}^p$, the other around $-\boldsymbol{\mu}$. Most of (low-dimensional) machine learning algorithms are anchored in this mentally convenient visualization. But the large-dimensional image is completely different. Standard Gaussian vectors $\mathbf{x} \in \mathbb{R}^p$ have an Euclidean norm of order $\|\mathbf{x}\| \sim O(\sqrt{p})$ but a spread of order $\|\mathbf{x}\| - \mathbb{E}[\|\mathbf{x}\|] \sim O(1)$, and nontrivial classification can be performed as long as $\|\boldsymbol{\mu}\|$ is no smaller than order $O(1)$. The mental image is thus one of two spheres in $\mathbb{R}^p$ with an extremely large radius (of order $O(\sqrt{p})$), around which the data of both classes "accumulate." Figure 1.5 provides a comparative picture for small- versus large-dimensional classification.

With this image in mind, the Euclidean distance paradigm is shifted: For small p, the information lies in the typical distance from one data point to a "centroid"; for large p, the centroid is far from *all* data points (it lives in an "empty" region of the space), and the class information is summarized in the accumulated small, deterministic deviations of all data points from the same class; this deviation is (asymptotically) invisible for any data vector but can be inferred collectively from the large data matrix.

[7] The *statistical* information contained in the data such as the mean $\mathbb{E}[\mathbf{x}] \in \mathbb{R}^p$ can be sparse (i.e., has a few nonzero entries), but the practical large-dimensional data vectors must randomly "fluctuate" with sufficiently many degrees of freedom around their possibly low-dimensional manifold structure. The large-dimensional random fluctuation of the data is essential to produce a statistically "robust" behavior of the algorithms and is key to establishing mathematical convergence in the large n, p setting.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

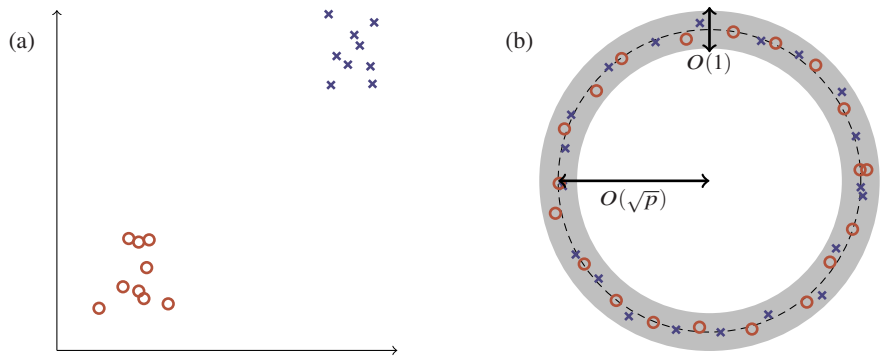


Figure 1.5 Visual representation of classification in **(a)** small and **(b)** large dimensions. The **red** circles and **blue** crosses represent data points from different classes.

Consequently, machine learning algorithms based on the evaluations of Euclidean distances $\|\mathbf{x}_i - \mathbf{x}_j\|$, inner products $\mathbf{x}_i^\mathsf{T}\mathbf{x}_j$, nonlinear activations $\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_i)$, regressions $f(\beta^\mathsf{T}\mathbf{x}_i)$, etc., of data $\mathbf{x}_i$ or data pairs $\mathbf{x}_i,\mathbf{x}_j$ structurally behave differently in large dimensions (from their small-dimensional counterparts).

Core Random Matrices in Machine Learning Algorithms

Be it in a supervised, semi-supervised, or unsupervised context, machine learning algorithms essentially consist of extracting structural information from some available set of data $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathcal{X}$: this is done, in general, via one-to-one comparisons of the data. At the heart of most algorithms, we notably find affinity matrices of the type:

$$\mathbf{K} \equiv \left\{\kappa(\mathbf{x}_i,\mathbf{x}_j)\right\}_{i,j=1}^{n} \in \mathbb{R}^{n\times n}, \tag{1.12}$$

where $\kappa\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ evaluates the closeness or affinity between $\mathbf{x}_i$ and $\mathbf{x}_j$. For graphs, the data $\mathbf{x}_i$ are merely the nodes (or vertices) of the graph, and $\kappa(\mathbf{x}_i,\mathbf{x}_j) = w_{ij}$ is thus the weight of the edge (i,j), which may be real of binary (i.e., $w_{ij} \in \{0,1\}$ depending on whether node i attaches to node j).

For $\mathcal{X} = \mathbb{R}^p$ and $\mathbf{x}_i$ statistically distributed, this naturally gives rise to a family of *kernel random matrices*, among which are inner-product kernel random matrices with $\kappa(\mathbf{x}_i,\mathbf{x}_j) = f(\mathbf{x}_i^\mathsf{T}\mathbf{x}_j)$, distance-based kernel random matrices with $\kappa(\mathbf{x}_i,\mathbf{x}_j) = f(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, and correlation random matrices with $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^\mathsf{T}\mathbf{x}_j/(\|\mathbf{x}_i\|\cdot\|\mathbf{x}_j\|)$. In the first case, f is often taken to be either linear $f(t) = t$ (therefore giving rise to sample covariance or Gram matrix models), a polynomial $f(t) = a_k t^k + \ldots + a_0$, or of a sigmoid type, such as the logistic function $f(t) = (1+e^{-x})^{-1}$ or the hyperbolic tangent $f(t) = \tanh(t)$. In the second case, f can be either linear (and we obtain a Euclidean distance matrix [Dokmanic et al., 2015]) or, more often, $f(t) = \exp(-t/(2\sigma^2))$ for some $\sigma > 0$, which is referred to as the *heat* kernel, the *Gaussian* kernel, or the RBF kernel.

When the $\mathbf{x}_i$s themselves are not directly separable in their ambient space, they are conventionally mapped into a *feature space*, in which they become separable. As feature extraction is possibly the single most important but usually hardest task in machine learning, it comes in a variety of forms. Kernel matrices of the type (1.12)

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

typically play the role of a feature extraction method, which maps the data points into a *reproducing kernel Hilbert space* (RKHS) [Schölkopf and Smola, 2018]. Another closely related, yet equally popular, approach is random extraction by means of *random feature maps*, which consist in operating $\sigma(\mathbf{W}\mathbf{x})$ for some (usually randomly and independently drawn) matrix $\mathbf{W} \in \mathbb{R}^{N\times p}$ and some nonlinear function $\sigma\colon \mathbb{R}^N \to \mathbb{R}^N$ applying entrywise, i.e., $\sigma(\mathbf{y}) = [\sigma_0(y_1),\ldots,\sigma_0(y_N)]^\mathsf{T}$ for some $\sigma_0\colon \mathbb{R} \to \mathbb{R}$, which, with a slight abuse of notation, we simply call σ. Among random feature maps, the most popular is the *random Fourier features* method proposed by Rahimi and Recht [2008], for which $\sigma(t) = \exp(-\imath t)$ (so, formally, $\sigma(\mathbb{R}) \subset \mathbb{C}$ rather than $\mathbb{R}$ in this case).

Neural networks operate likewise. Every size-N layer (that contains N neurons) of a neural network operates $\sigma(\mathbf{W}\mathbf{x})$ for an input $\mathbf{x}$, a linear mapping $\mathbf{W} \in \mathbb{R}^{N\times p}$ (the neural weights to be learned), and a nonlinear *activation function* $\sigma\colon \mathbb{R} \to \mathbb{R}$.[8] In this setting, σ is usually taken to be a sigmoid function (the logistic function, the tanh, or the Gaussian error function), or, more recently, the rectified linear unit (ReLU) function $\sigma(t) = \max(0,t)$.

Collecting the data in $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n] \in \mathbb{R}^{p\times n}$, the sample covariance matrix of the random features of the data then reduces to the Gram matrix:

$$\mathbf{\Phi} \equiv \sigma(\mathbf{W}\mathbf{X})^\mathsf{T}\sigma(\mathbf{W}\mathbf{X}), \tag{1.13}$$

which is also a central object of interest in this book.

The aforementioned kernel and Gram matrices of feature maps are actually much interrelated. For instance, the random Fourier features $\sigma(\mathbf{W}\mathbf{x})$, with $\sigma(t) = \exp(-\imath t)$ and $\mathbf{W} \in \mathbb{R}^{N\times p}$ having i.i.d. standard Gaussian entries, that is, $\mathbf{W}_{ij} \sim \mathcal{N}(0,1)$, are known to have the fundamental property:

$$\frac{1}{N}\mathbb{E}_{\mathbf{W}}[\sigma(\mathbf{W}\mathbf{x})^\mathsf{T}\sigma(\mathbf{W}\mathbf{y})] \equiv \exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2\right),$$

so that random Fourier features are intricately connected to Gaussian kernel matrices. This property ensures, in particular, that the Gaussian kernel $\kappa(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/2)$ is a *nonnegative definite kernel* in the sense that $\mathbf{K} = \{\kappa(\mathbf{x}_i,\mathbf{x}_j)\}_{i,j=1}^n$ is a nonnegative definite matrix (for any n and any set of $\mathbf{x}_1,\ldots,\mathbf{x}_n$), a particularly convenient property in both theoretical and practical kernel learning. An important subclass of kernel functions, referred to as Mercer kernels [Schölkopf and Smola, 2018], share this nonnegative definiteness property and have long been privileged in machine learning. We shall see in this book that, from a large-dimensional perspective, Mercer kernels can be, in general, suboptimal, and that simple but less intuitive choices of κ can largely outperform these conventional kernels.

A large body of machine learning algorithms (spectral clustering, linear, or logistic regression, support vector machines, and neural networks) relates, in one way or another, to the aforementioned *global properties* (eigenvalues, content of dominant

[8] Sometimes, an additional *bias* term is considered and the network operates $\sigma(\mathbf{W}\mathbf{X}) + \mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^N$ also to be learned.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

eigenvectors, linear, or nonlinear functionals of the resolvent) of the above matrices $\mathbf{K}$ or Φ. A systematic statistical analysis of these global properties for all finite p,n,N is, however, often out of reach, even for the simplest standard Gaussian modeling of the data.

In this book, we will show that random matrix theory manages to leverage the large-dimensional nature of both the data and the learning systems (i.e., large n,p,N), to tackle this statistical analysis. We will see, in particular, that several conventional models for $\mathbf{K}$ can be "Taylor-expanded" under the form of matrices involving only first- and second-order moments of the data distribution. The Gram matrix Φ cannot be directly Taylor-expanded in this way (it will be "Hermite-polynomially expanded" though) but will also be shown to behave as a kernel random matrix and be decomposed as the sum of more elementary random matrices, the statistical properties of which also become tractable in the large-dimensional regime.

In short, the intractable matrices $\mathbf{K}$ and Φ will be approximated by tractable ersatz $\tilde{\mathbf{K}}$ and $\tilde{\Phi}$, which behave asymptotically the same in the sense that

$$\|\mathbf{K} - \tilde{\mathbf{K}}\| \xrightarrow{\text{a.s.}} 0, \quad \|\Phi - \tilde{\Phi}\| \xrightarrow{\text{a.s.}} 0,$$

in *operator norm* as $n,p,N \to \infty$ at a similar rate. These matrices $\tilde{\mathbf{K}}$ and $\tilde{\Phi}$ will allow for further and deeper mathematical analysis.

Performance Analysis: Spectral Properties and Functionals

In a classification context, where, conventionally, $\mathbf{x}_i \in \mathbb{R}^p$ belongs to one of the k classes $\mathcal{C}_1,\ldots,\mathcal{C}_k$ with $k \ll n$ (the number of data samples), and thus $k \ll p$ whenever $p \sim n$, the approximation matrices $\tilde{\mathbf{K}}$ and $\tilde{\Phi}$ will often be shown to take a *spiked random matrix* form. That is, for instance,

$$\tilde{\mathbf{K}} = \mathbf{Z} + \mathbf{P},$$

where $\mathbf{Z} \in \mathbb{R}^{n \times n}$ is a random symmetric matrix, in general, having entries of zero mean and rather "uniform" variances, while $\mathbf{P} \in \mathbb{R}^{n \times n}$ is a *low-rank* matrix (the rank of which is often related to k), comprising the statistical information about the data-class associations and the statistical properties of the classes.

These spiked random matrix models have been extensively studied, and it is possible to extract much information about them. In particular, the dominant eigenvectors of $\tilde{\mathbf{K}}$ are known to relate to the eigenvectors of $\mathbf{P}$ (which carry the sought-for data-class information) whenever a *phase transition* threshold is exceeded.

In a regression setting where the $\mathbf{x}_i$s are assumed independently and identically distributed, the regression vector $\boldsymbol{\beta}$ of interest is a certain functional of $\mathbf{K}$ or Φ. For instance, a random feature regression from the observations $\mathbf{X} \in \mathbb{R}^{p \times n}$ to the desired outputs $\mathbf{y} \in \mathbb{R}^n$ entails the regression vector:

$$\boldsymbol{\beta} = \sigma(\mathbf{W}\mathbf{X}) \left(\Phi + \gamma \mathbf{I}_n\right)^{-1} \mathbf{y},$$

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

which is thus an (indirect) function of the resolvent $\mathbf{Q}_\Phi(-\gamma) = (\Phi + \gamma \mathbf{I}_n)^{-1}$ of Φ for a certain $\gamma > 0$. Random matrix theory possesses tools to analyze the statistical properties of such vectors $\boldsymbol{\beta}$ as well.

Least squares support vector machines and most conventional algorithms of graph-based semi-supervised learning relate to functionals of the same type. This also holds true (yet less directly) for nonlinear (e.g., logistic) regression, where $\boldsymbol{\beta}$ is *implicitly defined* as a function of $\mathbf{Q}_\Phi$. Similarly, in their plain form, support vector machines can be seen as nonlinear regressors which also fall within this scope.

Since eigenvalues, eigenvectors, and regressor statistics are at the core of machine learning algorithm performance, once these central quantities are accessible, the actual (asymptotic) classification error rates, mean squared error of regression, etc., become also accessible. It is important to point out here that not only bounds on performance but *actual accurate estimators* of the performance are provided. Under a random matrix framework, a *precise* characterization of the anticipated performance (as well as its error margins) for the above algorithms becomes available.

Since these performance indicators depend on the various hyperparameters of the problem, themselves being quantifiable from data statistics, in many scenarios, it becomes possible to fine-tune the algorithms without resorting to cross-validation procedures. We shall notably see how some simple instances of neural networks can be fairly well understood: why the rectifier $\max(t,0)$ is a convenient choice, and how the activation function and the data statistics mix up, etc. We will also understand that kernel methods do *not* function as one may think they should, and that there exists an elegant interplay between data statistics and the successive derivatives of the kernel function at a precise position.

Directions of Improvement and New Ideas

Due to the complete change of paradigm when comparing data from a small-versus a large-dimensional perspective, the overall behavior and the ensuing performance of the studied algorithms are often tainted, when large-dimensional data are handled.

We shall notably see, in the course of the book, that the conventional heat (or Gaussian) kernel used in various classification contexts is largely suboptimal. We shall also see that most graph-inspired semi-supervised learning algorithms in the literature *fail* to properly accomplish their requested task for n,p large and comparable; yet, we will show that the so-called PageRank approach [Avrachenkov et al., 2012] happens not to fail, although the fundamental reasons behind its nondegrading performance are at odds with the initial inspiration for the method; but most importantly, this popular approach will also be shown to perform quite far from optimal and, in particular, *not* to be capable of benefiting from a large addition of unlabeled data. This observation entails the very unpleasant property that purely unsupervised methods tend to outperform semi-supervised ones when the number of unlabeled data is large.

For all these applications, the book will list a set of recommendations and improved methods, which are tailored to large (as well as practically not so large)-dimensional data learning. Among others, optimal, but quite counterintuitive, kernel functions

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

will be introduced, new regularization procedures for supervised and semi-supervised learning will be discussed that particularly defeat the "curse of dimensionality" in semi-supervised learning (by fully exploiting the additional information from unlabeled data), and some further light on the design of neural networks will be cast.

1.2.4 Exploiting Universality: From Large-Dimensional Gaussian Vectors to Real Data

Before delving into the core of the manuscript, we conclude this section by further elaborating on the universality phenomenon briefly discussed above, which is of much greater importance to machine learning than one may anticipate.

First, let us recall that most random matrix results derived in the literature, even the most recent ones on machine learning applications (to be discussed in this book), are based on the assumption of data either arising from (possibly a mixture of) Gaussian distributions or represented by random vectors with independent entries. These models are generally deemed unsuitable to mimic real data, and we will *not* claim otherwise. It is a fact that real data, such as images, are largely more complex than mere Gaussian vectors.

Yet, what we do claim here and throughout this book is that *scalar observations* (regressor or classifier outputs, misclassification rates, etc.) obtained from large-dimensional and numerous data *tend to behave as if the data were Gaussian* (mixtures) in the first place. This is a fundamental disruption from small-dimensional statistics that random matrix analysis structurally exploits: Rather than assuming data as fixed entities living in a complex manifold, random matrix theory mostly exploits their numerous degrees of freedom, which, by universality, induce deterministic behavior in the large-dimensional limit, thus *independently* of the underlying vector data distribution.

We justify this claim below with both empirical and theoretical arguments.

Theory versus Practice

Our first argument follows after numerous comparative experiments made between theoretical findings on Gaussian versus real data. Indeed, although mostly derived under simple and seemingly unrealistic Gaussian mixture models, many theoretical results mentioned above show an *unexpected close match* when applied to popular real-world (sometimes not so) large-dimensional datasets, such as the MNIST handwritten-digit dataset [LeCun et al., 1998], the related Fashion-MNIST [Xiao et al., 2017], Kannada-MNIST [Prabhu, 2019] and Kuzushiji-MNIST [Clanuwat et al., 2018] datasets, the German Traffic Sign dataset [Houben et al., 2013], deep neural network features of the now popular ImageNet dataset [Deng et al., 2009], used for state-of-the-art machine learning and computer vision applications, as well as numerous financial and electroencephalography (EEG) time series datasets. In particular, while most elementary machine learning methods discussed in this book cannot be applied directly on raw ImageNet images to yield satisfactory performance, when performed on "deep" features of the data (such as VGG, DenseNet, or ResNet features) obtained from *independent* deep neural networks, these algorithms tend to behave the

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

same as with simple Gaussian mixtures [Seddik et al., 2020]. These seemingly striking empirical observations are indeed theoretically sustained by universality arguments arising from the powerful concentration of measure theory.

To be more precise, the following systematic comparison approach will be pursued in this book. An *asymptotically nontrivial* classification or regression problem is studied: that is, we assume that the problem at hand is theoretically neither too easy nor too hard to solve (as the one discussed in Section 1.1.3) and practically leads, in general, to, say, (binary) classification error rates of the order of $5\%-30\%$ and of relative regression errors also of the order $5\%-30\%$. In particular, we insist that the asymptotic random matrix framework under study is, in general, incapable to thinly grasp error rates below the 1%–2% region, which may be the domain of "outliers" and marginal data.

Having posed this nontriviality assumption, we shall generically model the data as being drawn from a simple mixture model, for example, the Gaussian mixture model that gives access to a large panoply of powerful technical tools. The theoretical results obtained from the proposed analyses (asymptotic performance notably) are thus function of the statistical means and covariances of the mixture distribution. To compare the theoretical results to real data, we then conduct the following procedure:

(i) exploiting the numerous and labeled samples of the real datasets (such as the $\sim$60 000 images of the training MNIST database), we empirically estimate the *scalar* functions of the statistical means and covariances (that determine the asymptotic performance of the method under study), for each class in the database;
(ii) we then evaluate the asymptotic performance that a genuine Gaussian mixture model having *these means and covariances* would have;
(iii) we compare these "theoretical" values to actual simulations.

As the book will demonstrate in most scenarios, this procedure systematically leads to the conclusion that the *performance of machine learning methods obtained on mere Gaussian mixtures* approximate surprisingly well the performance observed on real data and features. On a side note, we mentioned in Remark 1.2 that it is likely inappropriate to use the sample covariance matrix to estimate the population covariance of the small (i.e., n not much larger than p) databases, such as the MNIST database (for which $n/p \ll 100$). However, it turns out that, as the quantities of interest (e.g., classification or regression errors) are generally *scalar* functionals of the data statistical means and covariances, it is still possible, in the large n,p regime, to derive *consistent* estimators of these quantities without resorting to an exact evaluation of the (large-dimensional) moments; see more discussions on this topic in Sections 3.2 and 4.4.

As already mentioned in Remark 1.3, this surprising accordance between theory and practice is possibly due to the *universality* of random matrix results, that is, only the first several order statistics of the data/features at hand matter in the large-dimensional regime (recall for instance that the limiting eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\mathbf{X} \in \mathbb{R}^{p\times n}$ having i.i.d. zero mean and unit variance entries is the *same* Marčenko–Pastur law, irrespective of the higher order moments of $\mathbf{X}$).

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

Yet, another stronger argument can be made, especially when it comes to machine learning for image processing.

Concentrated Random Vectors and Real Data Modeling

The modeling assumption that the data vectors $\mathbf{x}_i$ are linear or affine maps $\mathbf{x}_i = \mathbf{A}\mathbf{z}_i + \mathbf{b}$ of random vectors $\mathbf{z}_i$ constituted of i.i.d. entries is simultaneously an asset for random matrix analysis (by exploiting the degrees of freedom in the entries of $\mathbf{z}_i$) but a severe practical limitation, as few real datasets are likely of this simplistic form.

El Karoui [2009] provided a first means for random matrix theory to go beyond the "vector of independent entries" assumption.[9] There, relying on elements of the *concentration of measure theory*, extensively developed by Ledoux [2005], El Karoui essentially shows (in a rather technical manner) that some of the early random matrix results from Pastur, Bai, and Silverstein remain valid under the assumption that the $\mathbf{x}_i$s are *concentrated random vectors*. Roughly speaking, a random vector $\mathbf{x} \in \mathbb{R}^p$ is *concentrated* if, for a certain family of functions $f : \mathbb{R}^p \to \mathbb{R}$, there exists a deterministic scalar $M_f \in \mathbb{R}$ such that

$$\mathbb{P}\left(|f(\mathbf{x}) - M_f| > t\right) \leq \alpha(t) \tag{1.14}$$

for some decreasing function $\alpha : \mathbb{R} \to \mathbb{R}$; in general, $\alpha(t)$ will be of the form $\alpha(t) = Ce^{-ct^q}$ for some $q > 0$ and $C, c > 0$ constants (which may depend on p though). Intuitively, a concentrated random vector is a (random) point in high-dimensional space having "predictable *scalar* observation" $f(\mathbf{x})$, in the sense that, with (exponentially) high probability, $f(\mathbf{x})$ takes values very close to the deterministic M_f. Thus, in the (one-dimensional) "observable world," the observation $f(\mathbf{x})$, which may typically be any performance metric of a machine learning algorithm on a test datum $\mathbf{x}$, appears to be "stable" for any concentrated vector $\mathbf{x}$.[10]

Ledoux and El Karoui mostly focused on concentrated random vectors defined on Lipschitz classes of functions f, that is, $\mathbf{x}$ is *Lipschitz-concentrated* if (1.14) holds for all f such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. These stringent constraints, however, make it hard to find random vector belonging to this class. As a matter of fact, in this class, the only standard random vectors are the Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and the uniform vector on the sphere $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\| \sim \mathbb{S}^{p-1}$ for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. However, quite importantly, *every* $\mathbb{R}^p \to \mathbb{R}^q$ *Lipschitz-mapping* $g(\mathbf{x})$ *and* $g(\mathbf{u})$ of these two random vectors, by definition, also belong to the class.[11]

A visual representation of the notion of concentration is presented in Figure 1.6.

Yet, since the widest class of (Lipschitz) concentrated random vectors is restricted to Lipschitz maps of standard Gaussian vectors, at first sight, concentrated random

[9] See also Pajor and Pastur [2009] published in the same year under slightly more constrained assumptions.

[10] Note that by modeling the input data $\mathbf{x}$ as a concentrated random vector and stating that the output (statistics) of a machine learning algorithm is "stable" implicitly assumes some *regularity* in the algorithm, which, as we shall see, can be shown to hold for many popular methods including deep neural networks (and which often takes the form of a "Lipschitz control").

[11] Under the more restricted class of *Lipschitz and convex* functions, random vectors with i.i.d. and bounded entries (up to normalization) also create a class of (convexly) concentrated random vectors.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

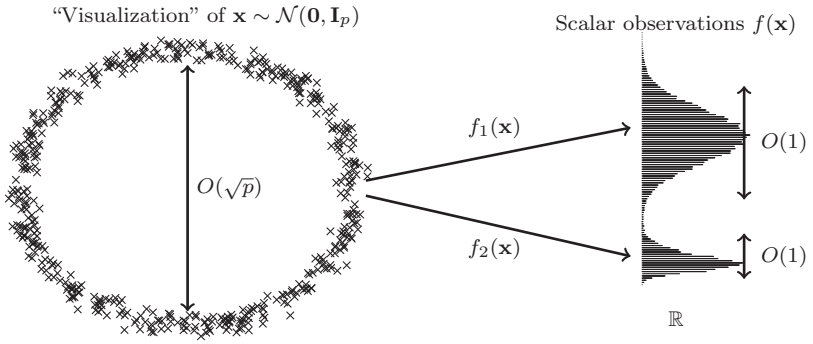


Figure 1.6 Multivariate Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, a fundamental example of concentrated random vectors. (**Left**) A visual "interpretation" of 500 independent drawings of $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. (**Right**) Concentration of observations for linear ($f_1(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_p/\sqrt{p}$) and Lipschitz ($f_2(\mathbf{x}) = \|\mathbf{x}\|_\infty$) maps.

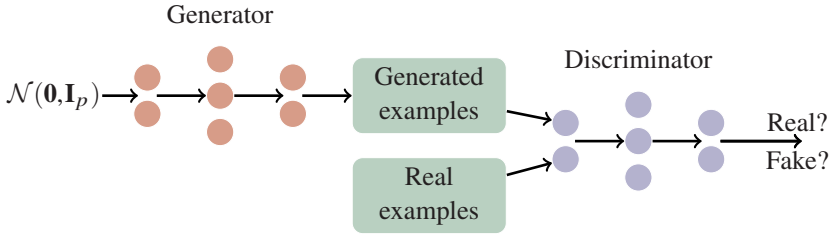


Figure 1.7 Illustration of a generative adversarial network (GAN).

vectors are seemingly no more elaborate models than linear and affine maps of Gaussian vectors. As a consequence, there is a priori no reason to assume that the mixtures of concentrated random vectors can model real data any better than Gaussian mixtures.

It turns out that this intuition is again tainted by erroneous small-dimensional insights. Indeed, there *practically exist extremely data-realistic concentrated random vectors*: the outputs of GANs [Goodfellow et al., 2014], as shown in Figure 1.7. GANs generate artificial images $g(\mathbf{x})$ from large-dimensional standard Gaussian vectors $\mathbf{x}$, where g is a conventional feedforward neural network trained to mimic real data. As such, g is the combination of Lipschitz nonlinear (the neural activations) and linear (the inter-layer connections) maps, and is thus a Lipschitz mapping.[12] The output image vectors $g(\mathbf{x})$, see examples in Figure 1.8, are thus concentrated vectors. Modern GANs are so sophisticated, that it has become virtually impossible for human beings to tell whether their outputs are genuine or artificial. This, as a result, strongly suggests that concentrated random vectors are accurate models of real-world data.

[12] In practice, other operations are also performed in neural networks, such as pooling operations, random or deterministic dropouts, and various connectivity matrix normalization procedures, so as to achieve better performance. They are all shown to be Lipschitz [Seddik et al., 2020].

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

Figure 1.8 Image samples generated by BigGAN in Brock et al. [2019].

A strong emphasis has thus lately been given to these models. The book will, in particular, elaborate on the work of Louart and Couillet [2018], which largely generalizes the seminal findings of El Karoui by providing a systematic methodological toolbox of *concentration theory for random matrices*. There, the notion of concentration is generalized by including *linear concentration*, which provides a consistent framework for the important notion of deterministic equivalents in random matrix theory, and by providing a wide range of properties and lemmas of immediate use for random matrix purposes.

An important finding of Louart and Couillet [2018] is that, first-order statistics of functionals of random matrices building from concentrated random vectors are *universal*; the asymptotic performance of many machine learning methods is, therefore, also universal. Specifically, for most conventional machine learning methods (support vector machines, semi-supervised learning, spectral clustering, random feature maps, linear regression, etc.), the asymptotic performance achieved on Gaussian mixtures $\mathcal{N}(\boldsymbol{\mu}_a, \mathbf{C}_a)$, $a \in \{1, \ldots, k\}$ coincides with that obtained on concentrated random vectors mixtures $\mathcal{L}_a(\boldsymbol{\mu}_a, \mathbf{C}_a)$, $a \in \{1, \ldots, k\}$, having the same means $\boldsymbol{\mu}_a$ and covariances $\mathbf{C}_a$ per class, and are *independent* of the high-order moments of the underlying distribution.

This strongly suggests that Gaussian mixture models, if not appropriate data "models" per se, are largely sufficient statistical assumptions for the theoretical understanding of real data machine learning.

Remark 1.4 (Concentration of measure, concentration inequalities, and nonasymptotic random matrices)**.** *It is important to raise here the fact that the concentration of measure theory is structurally broader than the scope of the popular* concentration inequalities *regularly used in statistical learning theory [Boucheron et al., 2013, Tropp, 2015, Vershynin, 2018]. Concentration inequalities are generally expressions of* (1.14) *for specific choices of f and their consequences, and they are, in particular, not new to random matrix theory. In Vershynin [2012] and Tao [2012], the authors exploit the mathematical strength of concentration inequalities (which, thanks to the exponential decay, is stronger and less cumbersome to handle than moment bounds) to prove fundamental results in random matrix theory. Yet, these inequalities are mostly exploited in proofs involving Gaussian or sub-Gaussian random vectors (as an instance of concentrated random vector). In particular, Vershynin establishes a nonasymptotic random matrix theory by exploiting concentration inequalities to bound various quantities of theoretical interest (notably bounds on the*

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

eigenvalue positions of random matrices). The book instead puts forth the interest of concentration of measure theory for data modeling *beyond a merely convenient mathematical tool.*

Concentration of measure theory is also all the more suited to machine learning as it structurally relates to linear, Lipschitz, or convex-Lipschitz functionals of random vectors and matrices. These are precisely the core elements of machine learning algorithms (kernels, activation functions, convex optimization schemes). From this viewpoint, concentration of measure theory is much more adapted to machine learning analysis than seemingly simpler data models. Note, for instance, that concentrated random vectors are stable (i.e., they remain concentrated) when passed through the layers of a neural network; this is particularly not true for Gaussian random vectors or vectors with independent entries, which, in general, no longer have independent entries when passed through nonlinear layers.

A last but not least convenient aspect of concentration of measure theory is that it flexibly allows one to "decouple" the behavior of the data size p and number n in the large-dimensional setting. It is technically much easier to keep track of *independent growth rates* for p and n under a concentration of measure framework than when exploiting more standard random matrix techniques (such as Gaussian tools to be discussed in Section 2.2.2).

1.3 Outline and Online Toolbox

1.3.1 Organization of the Book

The remainder of the book is divided into two parts.

Chapter 2 introduces the basics of random matrix theory *needed for machine learning applications in this book*. In doing so, we shall first revisit the traditional approach found in math-oriented sources, such as Bai and Silverstein [2010], based on a Stieltjes transform and truncation machinery, Pastur and Shcherbina [2011], based on a Gaussian-method approach, Tao [2012] and Vershynin [2012], based on concentration inequalities and a nonasymptotic random matrix approach, and also say a few words on Mingo and Speicher [2017], which follows a free probability framework and on Anderson et al. [2010], which is more oriented toward a determinantal point process and large deviations direction. Unlike most of these references though (with the possible exception of Pastur and Shcherbina [2011]), our methodology is primarily centered on the statistical analysis of the *resolvent* (and only secondarily on the Stieltjes transform) of random matrices, which is the chief object of interest to us in most machine learning applications. The particular mathematical toolbox exploited to derive the results is of secondary importance.

In this chapter, we will successively introduce:

- the fundamental notion of the *resolvent* $\mathbf{Q}(z) = (\mathbf{X} - z\mathbf{I}_n)^{-1}$ of a (random) matrix $\mathbf{X}$, and its relations to the eigenvalues of $\mathbf{X}$, the limiting spectrum of $\mathbf{X}$, the eigenvectors and eigenspaces associated with some specific eigenvalues, as well as

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

its relations to bilinear and quadratic forms often met in machine learning applications (linear or kernel regression, linear and quadratic discriminant analysis, support vector machines, as well as some simple neural networks);
- the almost equally important notion of *deterministic equivalents*, which extend the notion of the "limiting behavior" of large-dimensional random matrices, when such limits may not exist (which is the case of most structured random matrix models of practical interest); *deterministic equivalents for the resolvent* of random matrix models are at the core of almost all results derived in this book;
- the foundational Marčenko–Pastur and Wigner semicircle laws, which, as we shall see, serve as a reference "null model" to all random matrix models met in machine learning applications; even quite sophisticated random matrix transformations (through nonlinear kernels, and activation functions, etc.) will be seen to boil down, in one way or another, to either one (or a mixture of both) of these reference laws;
- a successive presentation of the three main technical tools at our disposal (in this book at least) to study random matrix models: the Bai–Silverstein Stieltjes transform approach, the Pastur–Shcherbina Gaussian tools, and the Louart–Couillet concentration of measure approach;
- the natural extensions of the Marčenko–Pastur- and Wigner-like random matrix models to more structured models: with correlation in either features or samples, with nonzero mean, divided into subclasses of correlated nonzero mean models, with a variance profile (in the case of heterogeneous graph models), etc.;
- a refined analysis of the large-dimensional spectrum of random matrices using tools from complex analysis, based on which statistical inference techniques on covariance matrix models are introduced;
- a thorough treatment of the so-called *spiked models* of random matrices, which carry a significant importance in the applications to machine learning: spiked models consist in *low-rank deviations* from some elementary or structured random matrix models; this "rank-sparsity" property simplifies the analyses and appropriately models the presence of cluster, classes, communities, principal components, etc., in machine learning problems;
- a short exposition of alternative tools and techniques, not of central focus in this book, but may have various advantages in specific random matrix structures;
- a short presentation of the very recent concentration of measure theory for random matrices that extends most of the results presented in this chapter to much more realistic (generative) models of data for machine learning applications.

This lengthy chapter provides a vast majority of the necessary tools to conduct the analyses performed in the subsequent chapters of machine learning methods. This second "application" part is organized as follows:

- Chapter 3 introduces first applications of the proposed random matrix framework devised in Chapter 2 to detection, estimation, and statistical inference; particular emphasis is made on generalized likelihood ratio tests for the detection of information from noise, on linear and quadratic discriminant analysis in a binary

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

hypothesis test, on the estimation of distances between data statistics (particularly here, the estimation of distances between unknown covariances and divergences between Gaussian measures of unknown statistics), as well as on the performance of robust estimators of covariance (or scatter) matrices. The estimation of covariance distance is a typical example where the usual large-n alone statistical answer dramatically fails, even when the ratio n/p is quite large, and random matrix analysis provides consistent (and improved) estimators. As for robust M-estimators, it is typical of a scenario where classical statistics fail to perform any satisfying analysis, while random matrix methods exploit concentration of measure phenomena to fully understand and improve their behavior.

- Chapter 4 follows with a detailed exposition of kernel random matrices and their applications to kernel- and graph-based methods in machine learning. This chapter successively exposes the many consequences for these methods of the already several times discussed *concentration of distances* phenomenon and shows that, as a result, the behavior, performance, and the role of hyperparameters (kernel function, regularization penalty, etc.) become tractable and amenable to improvement. Applications to kernel spectral clustering, graph-based semi-supervised learning (SSL), and kernel ridge-regression (also referred to as least-squares support vector machine [SVM]) are investigated, as representative examples of unsupervised, semi-supervised, and supervised learning methods. All these methods will be shown to be theoretically tractable, easy to optimize and thus to improve, with experiments on real data confirming the theoretical findings. The specific example of SSL is quite telling of the limitations of standard small-dimensional intuitions, and it will be shown that *all existing* classical graph-based SSL methods either dramatically fail, or, at best, do not exhibit the expected SSL behavior (notably failing to account for the large number of unlabeled data): The proposed random matrix approach is quite simple and is proven to address this issue.
- Chapter 5 focuses specifically on neural network models. While modern deep neural networks remain hardly accessible, several studies are reported in this chapter that address simpler models of neural networks (with random and few layers, with a possibly recurrent structure) and for which, again, new insights and exact asymptotic performance are provided. An additional discussion of the learning dynamics of gradient descent methods is also exposed in which the step-by-step performance and the importance of early stopping mechanisms are theoretically analyzed.
- Chapter 6 goes a step beyond all previous chapters for which all metrics of interest (algorithm behavior, performance) are *explicit functions* of the random matrix models introduced in Chapter 2 (under the form of eigenvalue distribution, eigenvector statistics, bilinear forms on the resolvent, etc.): Here, we focus on convex-optimization schemes in machine learning having *no explicit solution*. As such, the performance of these algorithms is *implicitly* related to the random data matrix and seems, at first sight, not related to random matrix analysis. The chapter shows, instead, that most of these methods do exhibit asymptotic (large n,p)

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press

performance that can be expressed as an almost explicit function (via a few coupled equations) of random matrix models, thereby opening the door to a wide range of machine learning applications (logistic regression, support vector machines, general empirical risk minimization scheme, etc.).

- Chapter 7 discusses spectral methods for community detection on (mostly dense) graphs and networks. As opposed to all previous application chapters for which the elementary random matrix model under study is the Gram matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ for data matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$, the problem of community detection on graphs naturally relates to symmetric graph matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ with independent Bernoulli entries. The chapter discusses, at length, the popular stochastic block model (SBM) and degree-corrected SBM, which mimic, with a different degree of reality, the behavior of genuine graphs with communities. A short discussion on the (technically more challenging and so far not very random matrix-related) modern concern of community detection on the even more realistic case of large-dimensional and *sparse* graphs is also made.
- Chapter 8 closes the application chapters with a discussion on the extension of *all* aforementioned applications to real data modeling. There, using the recent concentration of measure for random matrix framework, simulations of extremely realistic models of data (images mostly) are used to validate the random matrix results devised in all previous chapters. The chapter notably conveys the fundamental but surprising message that simple data models (such as Gaussian mixtures) are often sufficiently rich to account for the large-dimensional behavior of many existing machine learning algorithms.

1.3.2 Online Codes

MATLAB as well as Python codes used to obtain most of the visual results (graphs, histograms) provided in the book are publicly available at https://github.com/Zhenyu-LIAO/RMT4ML.

https://doi.org/10.1017/9781009128490.002 Published online by Cambridge University Press