1

Introduction

1.1 Big Data: Basic Concepts

Data is "unreasonably effective" [2]. Nobel laureate Eugene Wigner referred to the unreasonable effectiveness of mathematics in the natural sciences [3]. What is big data? According to [4], its sizes are in the order of terabytes or petabytes; it is often online, and it is not available from a central source. It is diverse, may be loosely structured with a large percentage of data missing . It is heterogeneous.

The promise of data-driven decision-making is now broadly recognized [5–16]. There is no clear consensus about what big data is. In fact, there have been many controversial statements about big data, such as "Size is the only thing that matters."

Big data is a big deal [17]. The Big Data Research and Development Initiative has been launched by the US Federal government. "By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning" [17]. Universities are beginning to create new courses to prepare the next generation of "data scientists."

The age of big data has already arrived with global data doubling every two years. The utility industry is not the only one facing this issue (Wal-Mart has a million customer transactions a day) but utilities have been slower to respond to the data deluge. Scaling up the algorithms to massive datasets is a big challenge.

According to [18]:

> A key tenet of big data is that the world and the data that describe it are constantly changing and organizations that can recognize the changes and react quickly and intelligently will have the upper hand ... As the volume of data explodes, organizations will need analytic tools that are reliable, robust and capable of being automated. At the same time, the analytics, algorithms, and user interfaces they employ will need to facilitate interactions with the people who work with the tools.

1.1.1 Big Data—Big Picture

Data is a strategic resource, together with natural resources and human resources. Data is king! "Big data" refers to a technology phenomenon that has arisen since the late 1980s [19]. As computers have improved, their growing storage and processing capacities have provided new and powerful ways to gain insight into the world by sifting

Smart Grid using Big Data Analytics: A Random Matrix Theory Approach, First Edition.
Robert C. Qiu and Paul Antonik.

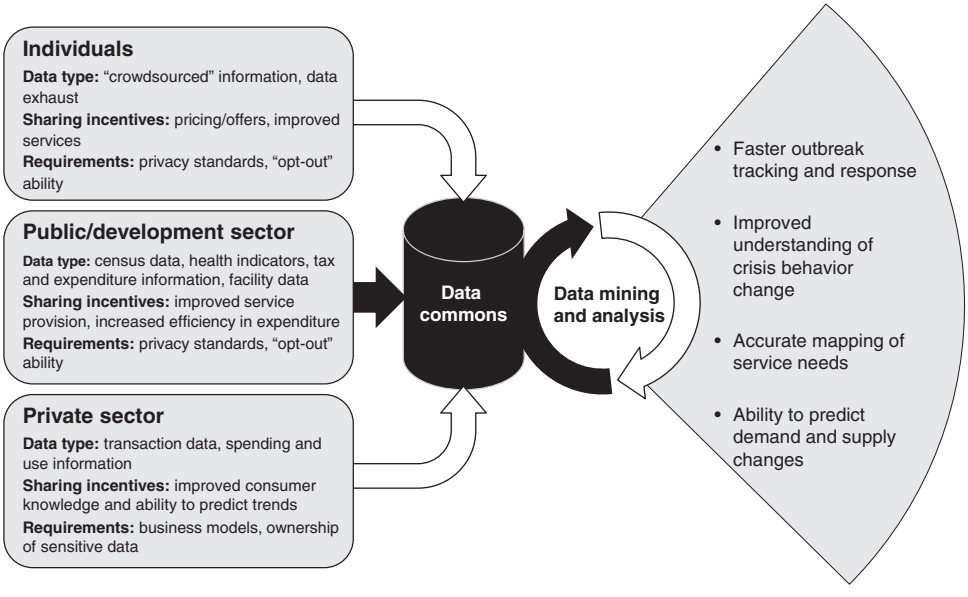

Figure 1.1 Big data, big impact: new possibilities for international development. Source: Reproduced from [6] with permission from the World Economic Forum.

through enormous quantities of data available. But this insight, discoverable in previously unseen patterns and trends within these phenomenally large data sets, can be hard to detect without new analytic tools that can comb through the information and highlight points of interest.

Sources such as online or mobile financial transactions, social media traffic, and GPS coordinates, now generate over 2.5 quintillion bytes of so-called "big data" every day. The growth of mobile data traffic from subscribers in emerging markets exceeded 100% annually through 2015. There are new possibilities for international development (see Figure 1.1).

Big data at the societal level provides a powerful microscope, together with social mining—the ability to discover knowledge from these data. Scientific research is being revolutionized by this, and policy making is next in line, because big data and social mining are providing novel means for measuring and monitoring wellbeing in our society more realistically, beyond the GDP, more precisely, continuously, everywhere [20].

Most scientific disciplines are finding the data deluge to be extremely challenging, and tremendous opportunities can be realized if we can better organize and access the data [16].

Chris Anderson believed that the data deluge makes the scientific method obsolete [21]. Petabytes data tell us to say correlation is enough. There is no need to find the models. Correction replaces causality. It remains open to see whether the data growth will lead to a fundamental change in scientific methods.

In the computing industry we are now focussing on how to process big data [22].

A fundamental question is "What is the unifying theory for big data?" This book adopts the viewpoint that big data is a new science of combining data science and information

science. Specialists in different fields deal with big data on their own, while information experts play a secondary role as assistants. In other words, most scientific problems are in the hands of specialists whereas only few problems—common to all fields—are refined by computing experts. When more and more problems are open, some unifying challenges common to all fields will arise. Big data from the Internet may receive more attention first. Big data from physical systems will become more and more important.

Big data will form a unique discipline that requires expertise from mathematics, statistics and computing algorithms.

Following the excellent review in [22], we highlight some challenges for big data:

- *Processing unstructured and semistructured data.* Presently 85% of the data are unstructured or semistructured. Traditional relational databases cannot handle these massive datasets. High scalability is the most important requirement for big-data analysis. MapReduce and Hadoop are two nonrelational data analysis technologies.
- *Novel approaches for data representation.* Current data representation cannot visually express the true essence of the data. If the raw data are labeled, the problem is much easier but customers do not approve of the labeling.
- *Data fusion.* The true value of big data cannot exhibit itself without data fusion. The data deluge on the Internet has something to do with data formats. One critical challenge is whether we can conveniently fuse the data from individuals, industry and government. It is preferable that data formats be platform free.
- *Redundancy reduction and high-efficiency, low-cost data storage.* Redundancy reduction is important for cost reduction.
- *Analytical tools and development environments that are suitable for a variety of fields.* Computing algorithm researchers and people from different disciplines are encouraged to work together closely as a team. There are enormous barriers for people from different disciplines to share data. Data collection, especially simultaneous collection for relational data, is still very challenging.
- *Novel approaches to save energy for data processing, data storage, and communication.*

1.1.2 DARPA's XDATA Program

The Defense Advanced Research Projects Agency's (DARPA's) XDATA program seeks to develop computational techniques and software tools for analyzing large volumes of data, both semistructured (e.g., tabular, relational, categorical, metadata) and unstructured (e.g., text documents, message traffic). Central challenges to be addressed include (i) developing scalable algorithms for processing imperfect data in distributed data stores, and (ii) creating effective human–computer interaction tools to facilitate rapidly customizable visual reasoning for diverse missions.

Data continues to be generated and digitally archived at increasing rates, resulting in vast databases available for search and analysis. Access to these databases has generated new insights through data-driven methods in the commercial, science, and computing sectors [23]. The defense section is "swimming in sensors and drowning in data." Big data arises from the Internet and the monitoring of industrial equipment. Sensor networks and the Internet of Things (IoT) are another two drivers.

There is a trend for data to be used that can sometimes be seen only once, for milliseconds, or can only be stored for a short time before being deleted, especially in some defense applications. This trend is accelerated by the proliferation of various digital devices and the Internet. It is important to develop fast, scalable, and efficient methods for processing and visualizing data.

The XDATA program's technology development is approached through four technical areas (TAs):

- TA1: Scalable analytics and data-processing technology;
- TA2: Visual user interface technology;
- TA3: Research software integration;
- TA4: Evaluation.

It is useful to consider distributed computing via architectures like MapReduce, and its open source implementation, Hadoop. Data collected by the Department of Defense (DoD) are particularly difficult to deal with, including missing data, missing connections between data, incomplete data, corrupted data, data of variable size and type, and so forth [23]. We need to develop analytical principles and implementations *scalable* to data volume and distributed computer architectures. The challenge for Technical Area 1 is how to enable systematic use of big data in the following list of topic areas:

- Methods for leveraging the problem structure to create new algorithms to achieve optimal tradeoffs among time complexity, space complexity, and stream complexity (i.e., how many passes over the data are needed).
- Methods for the propagation of uncertainty (i.e., every query should have an answer and an error bar), with performance guarantees for loss of precision due to approximations.
- Methods for measuring nonlinear relationships among data.
- Sampling and estimation techniques for distributed platforms, including compensating for missing information, corrupted information, and incomplete information.
- Methods for distributed dimensionality reduction, matrix factorization, matrix completion (within a distributed data store where data are not all in one place).
- Methods for operating on streaming data feeds.
- Methods for determining optimal cloud configurations and resource allocation with asymmetric components (e.g., many standard machines, a small number of large-memory machines, machines with graphical processing units).

The challenge for Technical Area 2 is how to hook up big data analytics to interfaces, including but not limited to the following topics:

- Visualization of data for scientific discovery, activity patterns, and summaries.
- Expressive visualization and/or query languages and processing that support domain-specific interaction, successive query refinement, repeated viewing of data, faceted search, multidimensional queries, and collaborative/interactive search.
- Principled design, including menus, query boxes, hover tips, invalid action notifications, layout logic, as well as processes of overview, zoom and filter, and details-on-demand.
- Support for the study and characterization of users, including extraction of relations and history, usage, hover time, click rate, dwell, etc.

- Functions of timeliness, online versus batch processing, metainformation, etc.
- Analytical workflows including data cleaning and intermediate processing.
- Tools for rapid domain-specific end-user customization.

1.1.3 National Science Foundation

The phrase "big data" in the National Science Foundation (NSF) refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, e-mail, video, click streams, and/or all other digital sources available today and in the future [5].

Today, US government agencies recognize that the scientific, biomedical and engineering research communities are undergoing a profound transformation with the use of large-scale, diverse, and high-resolution data sets that allow for data-intensive decision making, including clinical decision making, at a level never before imagined. New statistical and mathematical algorithms, prediction techniques, and modeling methods, as well as multidisciplinary approaches to data collection, data analysis and new technologies for sharing data and information are enabling a paradigm shift in scientific and biomedical investigation. Advances in machine learning, data mining, and visualization are enabling new ways of extracting useful information in a timely fashion from massive data sets, which complement and extend existing methods of hypothesis testing and statistical inference. As a result, a number of agencies are developing big-data strategies to align with their missions. The NSF's solicitation focuses on common interests in big data research across the National Institutes of Health (NIH) and the NSF.

1.1.4 Challenges and Opportunities with Big Data

There are challenges with Big Data. The first step is data acquisition. Some data sources, such as sensor networks, can produce staggering amounts of raw data. A lot of this data is not of interest. It can be filtered out and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information.

The second big challenge is to generate the right metadata automatically, and to describe what data is recorded and how it is recorded and measured. This metadata is likely to be crucial to downstream analysis. Frequently, the information collected will not be in a format ready for analysis. We have to deal with erroneous data: some news reports are inaccurate.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big data computing environments. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process, and bringing the data back.

Having the ability to analyze big data is of limited value if users cannot understand the analysis. Ultimately, a decision maker, provided with the result of analysis, has to interpret these results.

In short, there is a multistep pipeline required to extract value from data. This pipeline is not a simple linear flow—rather, there are frequent loops back as downstream steps suggest changes to upstream steps.

There has not been a commonly accepted definition of big data. In [24], there are some claims that may define the ballpark:

- Big data is the same as scalable analytics.
- Big-data problems are primarily on the application side.
- Big-data problems are primarily at the systems level.
- Big-data requires a cloud-based platform.
- The data-management community is in danger of missing the big-data train.
- It is not possible to conduct big-data research effectively without collaborating with people outside the data-management community.
- All the big-data problems can be reduced to MapReduce problems [25].
- The bulk of big-data challenges are being addressed by industry.
- The bulk of big-data challenges are at the implementation level.
- Size is the only thing that matters (for big data).

The growth of the data volume seems to outspend the advance of our computing infrastructure. Conventional data-processing technologies, such as database and data warehouse, are becoming inadequate for the amount of data.

1.1.5 Signal Processing and Systems Engineering for Big Data

The big-data workshop for signal processing and systems engineering was held in 2013 [4]. One motivation from the NSF's point of view [4] is to leverage analytical, computational, storage, and implementation tools:

- assess fundamental performance limits in processing and storage;
- develop scalable algorithms: online (adaptive) and decentralized;
- complement computer and information science and engineering (CISE) efforts on parallel architectures and computing;
- account for redundancy and error control: source and channel (de)coding;
- cross-fertilize NSF-wide advances on fault-tolerance, privacy, and security.

Another motivation is to facilitate ground-breaking research in big-data science and engineering:

- to offer top-down approaches for signal processing and systems engineering;
- to develop a toolbox for statistics and optimization.

High-level issues of interest include: Can lessons learned from "big systems" engineering be applied to big-data engineering? What are the right pathways? What are overarching tools to catalyze big-data collaboration between scientists and engineers? What are the grand challenges in big data science and engineering? How should we educate engineers about big data?

Big *engineering* data has unique characteristics: it is more disciplined and regulated. There are emerging engineering systems with big-data opportunities: smart grids, sensor nets, transportation, telemedicine, aerospace, testing, safety, nuclear, design blueprints and more. Now it may be necessary to rethink data collection and storage to facilitate big-data processing and inference tasks.

Some sample questions are: How do we trade off complexity for accuracy in massive decentralized signal and data analysis tasks? How can efficient signal and data analysis

algorithms be developed for big, unstructured or loosely structured data? What are the basic principles and useful methodologies to scale inference and learning algorithms and trade off the computational resources (e.g., time, space and energy) according to the needs of engineering practice (e.g. robustness versus efficiency, real time)?

Big-data processing and analysis, according to Hero [26], require the following: (i) Integration of very heterogenous data: correlation mining in massive database; processing data at vastly different scales and noise levels; processing a mixture of continuous and categorical variables. (ii) Reliable and robust quantitative models: uncertainty quantification; adaptation to drift over time. (iii) High throughput real-time processing: smart adaptive sampling and compression; distributed or parallel processing architectures. (iv) Interactive user interfaces: human-in-the-loop processing; visualization and dimensionality reduction.

Some signal-processing challenges, according to Hero [26], include the following. (i) Heterogeneous data integration: ranking signals for human-aided selection of relevant variables; fusing graphs, tensors, and sequence data; active visualization: dimensionality reduction. (ii) Flexible low-complexity modeling and computation: scalable signal processing: distributed algorithms and implementation; smart sampling: feedback-controlled signal search and acquisition. (iii) Reliable robust models for anomaly detection and classification: parsimonious signal processing; sparse correlation graphical models; decomposable signal processing: factored models and algorithms.

As for the signal-processing toolbox, we have the following primitives: linear equation solvers (Gauss, Givens, Householder); spectral representations (FFT, SVD); ensemble averaging (cross validation, bootstrap, boosting); optimization (linear least square, linear and quadratic programming, dynamic programing). They can used for the following applications. (i) Linear and nonlinear prediction: Wiener, Kalman, particle filtering, Volterra filters; (ii) signal reconstruction: matrix factorization, matrix, completion, robust principal component analysis (PCA). (iii) Dimension reduction: PCA, independent component analysis (ICA), independent principal component analysis (IPCA), canonical correlation analysis (CCA), linear discriminant analysis (LDA), nonlinear editing (NLE). (iv) Adaptive sampling: compressive sensing, distilled sensing, sketching. (v) Signal processing on graphs: graph spectra, the k-nearest-neighbor algorithm (k-NN) search, belief propagation.

There is a growing gap between the amount of data we generate and the amount of data we are able to store, communicate, and process. As Richard Baraniuk points out, we have produced already twice as much data as can be stored [27]. And the gap keeps widening. As long as this continues there is an urgent need for novel data-acquisition concepts like compressive sensing.

Compressive sensing and sparse representations play a key role: advanced probability theory and (in particular) random matrix theory, convex optimization, and applied harmonic analysis are becoming standard ingredients of the toolbox of many engineers. Compressive sensing has advanced the development of ℓ_1-minimization algorithms, and more generally of nonsmooth optimization. These algorithms find widespread use in many disciplines, including physics, biology, and economics [28]. The most important legacy of compressive sensing may be that it has forced us to think about information, complexity, hardware, and algorithms in a truly integrated manner.

Nondominated sorting is an interesting and useful framework for multicriteria anomalies, human-machine interaction, or multiple end users [29, 30].

The author's research proposals to the National Science Foundation (NSF) [31–34] are relevant in the context of this section.

1.1.6 Large Random Matrices for Big Data

Random matrices play a central role in statistics in the context of multivariate data. Three classical books are included here [35–37].

The continued growth of big data has given rise to high-dimensional statistical analysis. Convex analysis, Riemannian geometry and combinatorics are relevant. Random matrix theory (RMT) has emerged as a particularly useful framework for many theoretical questions associated with the analysis of high-dimensional multivariate data; see [38] for a recent overview of RMT.

RMT affects modern statistical thinking in two ways. On one hand, most of the mathematical treatments of RMT have focused on matrices with a high degree of independence in the entries, which one may refer to as "unstructured" random matrices. Recall that about 75% of big data is unstructured. On the other hand, in high-dimensional statistics, we are primarily interested in problems where there are lower dimensional structures buried under random noise.

In November 2011, the author of [39] dedicated over 200 pages to random matrix theory. In [40], the whole book was motivated by the same vision but with different regimes. The first book deals with so-called asymptotic regimes, while the second deals with nonasymptotic regimes. In the asymptotic regimes, the sizes of random matrices are assumed to approach infinity. For example, for a random matrix of $\mathbf{X}$ of size $m \times n$, we assume the asymptotic regime: $m \to \infty$, $n \to \infty$, but $m/n \to c$. On the other hand, the nonasymptotic regime is defined as: m and n are large, but *finite*. The author's research proposals to the National Science Foundation (NSF) [31–34] have a similar motivation.

As pointed out in Section 1.1.1, "High scalability is the most important requirement for big data analysis." Some state "Size is the only thing that matters." Based on this observation, it seems natural to the author to model big data using a nonasymptotic theory of random matrices. The motivation is to investigate how the algorithms *scale* with sizes of data samples.

We believe that a nonasymptotic theory of random matrices can unify many big-data problems. It is our intention to use this theory as the departure point for many problems studied later in this book.

Tensors (also known as multidimensional arrays or N-way arrays) are used in a variety of applications ranging from chemometrics to network analysis. The Tensor Toolbox [41] provides classes for manipulating dense, sparse, and structured tensors using MATLAB's object-oriented features.

1.1.7 Big Data Across the US Federal Government

We highlight some points [42] that which are relevant to the contex of this book.

The **Anomaly Detection at Multiple Scales** program at DARPA creates, adapts and applies technology to anomaly characterization and detection in massive data sets. Anomalies in data cue the collection of additional, actionable information in a wide variety of real-world contexts. The initial application domain is insider threat

detection in which malevolent (or possibly inadvertent) actions by a trusted individual are detected against a background of everyday network activity.

The Department of Energy (DOE) provides leadership to the data management, visualization and data analytics communities, including digital preservation and community access. **Mathematics for Analysis of Petascale Data** addresses the mathematical challenges of extracting insight from huge scientific datasets, finding key features and understanding the relationships between those features. Research areas include machine learning, real-time analysis of streaming data, stochastic nonlinear data-reduction techniques, and scalable statistical analysis techniques applicable to a broad range of DOE applications including sensor data from the electric grid, cosmology, and climate data.

The **Office of Basic Energy Sciences (BES) BES Scientific User Facilities** have supported a number of efforts aimed at assisting users with data management and analysis of big data, which can be as big as *terabytes* (1012 bytes) of data *per day* from a single experiment.

Researchers funded by the NSF are developing a unified theoretical framework for principled statistical approaches to network models with scalable algorithms in order to differentiate knowledge in a network from randomness.

Information Integration and Informatics funded by the NSF addresses the challenges and scalability problems involved in moving from traditional scientific research data to very large, heterogeneous data, such as the integration of new data types models and representations, as well as issues related to data path, information life-cycle management, and new platforms.

NSF funds a distinct discipline encompassing mathematical and statistical foundations and computational algorithms. High-speed networks distribute over 15 petabytes of data each year in real time from the Large Hadron Collider (LHC) at CERN in Switzerland to more than 100 computing facilities.

The **Theoretical and Computational Astrophysics Networks (TCAN)** program seeks to maximize the discovery potential of massive astronomical data sets by advancing the fundamental theoretical and computational approaches needed to interpret those data, uniting researchers in collaborative networks that cross institutional and geographical divides, and training the future theoretical and computational scientists.

There are research projects (i) developing data visualizations in the defense of massive computer networks, and (ii) transforming big data sets and big ideas about earth science theories into scientific discoveries.

1.2 Data Mining with Big Data

Big data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and data-collection capacity, big data is now rapidly expanding in all science and engineering domains, including physical, biological, and biomedical sciences.

Data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for big data applications is to explore the large volumes of

data and extract useful information or knowledge for future actions [43]. In many situations, the knowledge-extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. For example, the in-network processing in a large-scale cognitive radio network [44] is the bottleneck. For one microsecond of data collection, the processing time is in the level of several milliseconds (three orders of magnitudes larger). As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such big data.

Theorem 1.2.1 (HACE theorem [45]) Big data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data.

In the analogy of the blind men and the giant elephant, the localized (limited) view of each blind man leads to a biased conclusion. Exploring big data is equivalent to aggregating heterogeneous information from different sources (the blind men) to help draw a best possible picture to reveal the elephant in real time.

One of the fundamental characteristics of big data is the huge volume of data represented by heterogeneous and diverse dimensionalities. The reason is that different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations.

Autonomous data sources with distributed and decentralized controls are an important characteristic of big data applications. Being autonomous, each data source (a sensor) is able to generate and collect information without involving (or relying on) any centralized control.

While the volume of the big data increases, so do the complexity and the relationships underneath the data. One example is the time-varying wireless network or electric power grid.

A big-data processing framework is shown in Figure 1.2. The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because big data is often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For example, distributed detection and estimation [46] is relevant in the context of wireless sensor networks.

Example 1.2.2 (a long time series) We form a large random matrix using a long record of time series. Given a time series $x[i]$, $i = 1, \ldots, NT$, where N and T are integers, we form a large random matrix $\mathbf{X}$ of $N \times T$. For example, $N = 1000$ and $T = 4000$. We view the data as a number of data segments. Here we have N data segments; the length of each segment is T, so a total of NT data samples are needed. □

Example 1.2.3 (the square kilometer array (SKA)—a big-data viewpoint) The square kilometer array (SKA) (see Figure 1.3) has 2000–3000 dishes. The wavelength ranges from 3 m to 3 cm. The SKA will have an array of coherently connected antennas spread over an area about 3000 km in extent, with an aggregate antenna collecting area of up to 106 m^2 at centimeter and meter wavelengths. The project timeline has the telescope operational below 10 GHz by 2022.

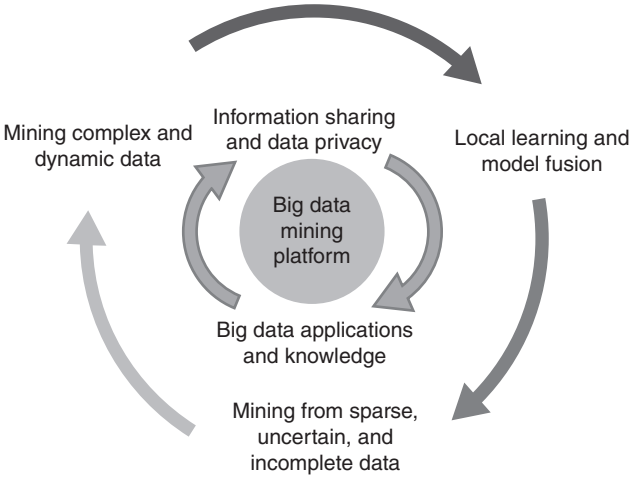

Figure 1.2 A big data processing framework. The research challenges form a three-tier structure and center around the "big data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high-level semantics, application-domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms. Source: Reproduced from [45] with permission.

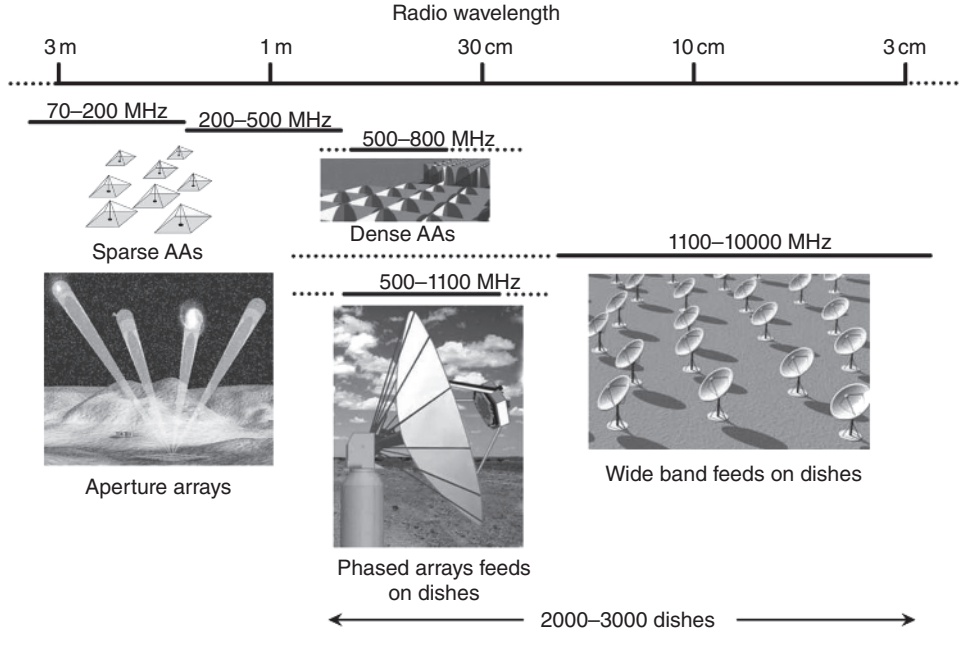

Figure 1.3 The square kilometer array. Source: Reproduced with permission from [47].

A large-scale wireless communication network is attempted to emulate a virtual array. So the SKA provides guidance.

With a 40 GB/s data volume, the data generated from the SKA are exceptionally large. We can model each dish as a sensor. So we deal with $N = 2000 - 3000$ sensors, which are spatially distributed. For each sensor, we observe a time series $\mathbf{x}_i \in \mathbb{C}^{T\times 1}$, for $i = 1, 2, \ldots, N$. We can collect the data from the N sensors into one single large matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}_{N\times T} \in \mathbb{C}^{N\times T}$$

The data for the SKA with time of T (called a snapshot) is represented by a large random matrix $\mathbf{X} \in \mathbb{C}^{N\times T}$. Now we study the time evolution of the data in a sequence of random matrices (for n snapshots) $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{C}^{N\times T}$. We can do some data processing using these large random matrices. (i) the sum of Hermitian random matrices (See Theorem 17.4.1) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbf{X}_1 \mathbf{X}_1^H + \ldots + \mathbf{X}_n \mathbf{X}_n^H \right)$ (ii) the product of non-Hermitian random matrices $\mathbf{X}_1 \cdots \mathbf{X}_n$; (iii) the geometric mean $\left(\mathbf{X}_1 \cdots \mathbf{X}_n \right)^{1/n}$. (iv) For N spatially distributed sensors (randomly), we form the data matrix $\mathbf{X}$ as above. What is the theoretical distribution of $\mathbf{X}$? It appears that this problem can be formulated in terms of a Euclidean random matrix. This problem corresponds to a random Green's function.

The so-called Euclidean random matrices, defined in Section 6.14, are a special class of random matrices. See also Section 16.1.5 for its connection with random geometric graphs. The elements A_{ij} of an $N \times N$ Euclidean random matrix $\mathbf{A}$ are given by a *deterministic* function f of positions of pairs of points that are randomly distributed in a finite region V of Euclidean space:

$$A_{ij} = f\left(\mathbf{r}_i, \mathbf{r}_j\right), \quad i, j = 1, \ldots, N$$

Here, the N points $\mathbf{r}_i$ are randomly distributed inside some region V of the d- dimensional Euclidean space with a uniform density $\rho = N/V$. □

Example 1.2.4 (local learning and model fusion for multiple information sources) As big data applications are featured with autonomous sources and decentralized controls, *aggregating distributed data sources* to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models, just like the elephant and blind men case. Large random matrices provide natural models for data representations in this context. We can form larger matrices using data matrices from distributed sources. The fundamental mathematical structure (random matrix) is kept invariant under the data fusion. The scalability, however, is relevant.

We can use the unifying tool of random matrix theory to study the resultant problem. The possibility of calculating eigenvalues without explicitly forming the sample covariance matrix allows us to study the problem in a distributed manner. See Section 16.3 for details.

Distributed estimation and detection is natural in this context. See Section 16.1 for details.

Model mining and correlations are the key steps. When the data is independent, identically distributed (i.i.d.)—noise only, the eigenvalue distribution has a rotational symmetry on the complex plane. When signal plus noise is present, some correlations are identified on the complex plane. Non-Hermitian random matrices are studied. This theory is a very recent breakthrough (Chapter 6). □

Example 1.2.5 (mining from sparse, uncertain, and incomplete data) Sparse, uncertain, and incomplete data are defining features for big data applications. For most machine-learning and data-mining algorithms, high-dimensional sparse data cause the reliability of the models derived from the data to deteriorate significantly. We must emphasize that sparsity and high dimensionality are two blessings, rather than curses, for data processing. The concentration of measurement phenomenon—unique to big data—can be exploited [40].

Uncertain data are a special type of data reality where each data field is no longer deterministic but is subject to some random/error distributions. This is mainly linked to domain-specific applications with inaccurate data readings and collection. In this book we promote the exploitation of randomness. Randomness is introduced as a natural resource for our use.

"Incomplete data" refers to missing data field values for some samples. The missing values can be caused by different factors, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values (e.g., dropping some sensor node readings to save power for transmission). Low-rank matrix recovery [40] deals with incomplete data. Again low-rank matrix recovery takes advantage of high-dimensionality of the data, by using large random matrices as the "sampling" matrix. □

Example 1.2.6 (mining complex and dynamic data) The rise of big data is driven by the rapid increasing of complex data and their changes in volumes and in nature [48]. Documents posted on WWW servers, Internet backbones, social networks, communication networks, and transportation networks, and so on are all featured with complex data. Simple data representations are insufficient. In big data, data types include structured data, unstructured data, semistructured data, and so on. Currently, there is no acknowledged effective and efficient data model to handle big data. In this book we pursue a paradigm of using large random matrices for data representations. This framework has the advantage of uncovering complex relationship networks in data. □

1.3 A Mathematical Introduction to Big Data

There is no standard definition for big data. We give a mathematical definition below.

Definition 1.3.1 (Fundamental Definition for Big Data) Big data must satisfy the following three conditions:

1. Data samples are modeled as random variables, say $X_1, X_2, \ldots, X_n$.

2. The number of data samples, say n, is sufficiently large that some limit results may be observed.
3. A function $f\left(X_1, \ldots, X_n\right)$ can be defined using n random variables.

The main motivation for this definition is to capture the mathematical implications of big data. In particular, we are interested in representing all the data samples in terms of a large random matrix $\mathbf{X}$; applications are modeled as the function $f(\mathbf{X})$.

Example 1.3.2 (data samples are independent random variables) In Definition 1.3.1, most of the time, we consider the special case of Condition 1 when the data samples are modeled by *independent* random variables. Combining Condition 1 with Condition 2, we can take advantage of a very large body of knowledge related to limit theorems in probability and statistics. Roughly speaking, when the size of independent random variables becomes large, some limits are approached.

The simplest and most thoroughly studied example is the sum of independent real-valued random variables. The key to the study of this case is summarized by the trivial but fundamental additive formulas

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right)$$

and

$$\psi_{\sum_{i=1}^{n} X_i}(\lambda) = \sum_{i=1}^{n} \psi_{X_i}(\lambda) \tag{1.1}$$

where $\psi_Y(\lambda) = \log \mathbb{E}e^{\lambda Y}$ denotes the logarithm of the moment-generating function of the random variable Y. $\mathbb{E}$ denotes the expectation. These formulas allow one to derive concentration inequalities of $Z = X_1 + X_2 + \cdots + X_n$ around its expectation via Markov's inequality. See [49].

If $X_1, \ldots, X_n$ are independent random variables taking values in $[a_1, b_1], \ldots, [a_n, b_n]$, the additivity formula (1.1) implies that

$$\psi_{Z-\mathbb{E}Z}(\lambda) \leqslant \frac{1}{2}\lambda^2 \nu \text{ for } \lambda \in \mathbb{R}$$

where $\nu = \sum_{i=1}^{n} \left(b_i - a_i\right)^2/4$. Since the right-hand side corresponds to the log-moment generating function of a centered normal random variable with variance ν, $Z - \mathbb{E}Z$ is said to be *sub-Gaussian* with variance factor ν. The sub-Gaussian property implies that $Z - \mathbb{E}Z$ has a sub-Gaussian tail. More precisely, we have, for all $t > 0$

$$\mathbb{P}\left\{[Z - \mathbb{E}Z] \geqslant t\right\} \leqslant 2\exp\left(-t^2/(2\nu)\right)$$

This is Hoeffding's inequality.

One of the simplest and more natural smoothness assumptions that one may consider is the so-called *bounded difference condition*. A function $f : \mathcal{X}^n \to \mathbb{R}$ of n variables

(all taking values in some measurable set $\mathcal{X}$) is said to satisfy the bounded differences condition if constants $c_1, \dots, c_n > 0$ exist such that for every $x_1, \dots, x_n, y_1, \dots, y_n \in \mathcal{X}^n$

$$\left| f\left(x_1, \dots, x_i, \dots, x_n\right) - f\left(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n\right) \right| \leqslant c_i$$

In other words, changing any of the n variables, while keeping the rest fixed, cannot cause a big change in the value of the function. Equivalently, one can interpret this as a Lipschitz condition.

The sum of bounded variables is the simplest example of a function of bounded differences. Indeed, if $X_1, \dots, X_n$ are real-valued independent random variables such that X_i takes its values in the interval $[a_i, b_i]$, then $f\left(X_1, \dots, X_n\right) = \sum_{i=1}^{n} X_i$ satisfies the bounded difference condition with $c_i = b_i - a_i$. The basic argument behind the martingale-based approach is that once the function satisfies the bounded difference condition, $Z = f\left(X_1, \dots, X_n\right)$ may be interpreted as a martingale with bounded increments with respect to Doob's filtration. In other words, we may write

$$Z - \mathbb{E}Z = \sum_{i=1}^{n} \Delta_i \tag{1.2}$$

where

$$\Delta_i = \mathbb{E}\left[Z \,|X_1, \dots, X_i\right] - \mathbb{E}\left[Z \,|X_1, \dots, X_{i-1}\right], \quad i = 1, \dots, n$$
$$\Delta_1 = \mathbb{E}\left[Z \,|X_1\right] - \mathbb{E}\left[Z\right].$$

The bounded difference condition implies that, conditionally on $X_1, \dots, X_{i-1}$, the martingale increment Δ_i takes it values in an interval of length at most c_i. Hence, Hoeffding's inequality remains valid for Z with $\nu = (1/4)\sum_{i=1}^{n} c_i^2$. This result is known as the bounded difference inequality, also often called *McDiarmid's inequality*. □

Example 1.3.3 (concentration inequalities for a nonasymptotic theory of independence) The study of random fluctuations of functions of *independent* random variables is the topic of concentration inequalities. Concentration inequalities quantify such statements, typically by bounding the probability that such a function is different from its expected value (or from its median) by more than a certain amount.

In the mid-1990s Michel Talagrand [50] provided major new insight: "a random variable that smoothly depends on the influence of many independent random variables satisfies Chernoff type bounds."

What kind of smooth conditions should we put on a function $f(\cdot)$ of independent random variables $X_1, \dots, X_n$ in order to get concentration bounds for $Z = f\left(X_1, \dots, X_n\right)$ around its mean or median?

One approach to understanding the concentration properties of Lipschitz functions of independent variables is based on investigating how product measures concentrate in *high-dimensional* spaces. The main ideas behind this approach are dominant in Talagrand's work.

In the above examples, we had only considered the linear combination $X_1, \ldots, X_n$ of independent random variables. Now we consider more general combinations $f(\mathbf{X})$ where we write $\mathbf{X} = (X_1, \ldots, X_n)$ for short.

The most powerful concentration of measure results, though, do not just exploit Lipschitz-type behavior in each individual random variable, but *joint* Lipschitz behavior.

One consequence of Talagrand's concentration theorem (Theorem 1.3.5) is the concentration of (empirical) spectral measure for a large random matrix [40].

We say the function $f : \mathbb{C}^n \to \mathbb{R}$ is a 1-Lipschitz function if $|f(\mathbf{x}) - f(\mathbf{y})| \leqslant \|\mathbf{x} - \mathbf{y}\|$ for all random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, where $\|\cdot\|$ is the Euclidean norm.

Theorem 1.3.4 (Gaussian concentration inequality for Lipschitz functions) Let $X_1, \ldots, X_n \equiv \mathcal{N}(0,1)$ be i.i.d. real Gaussian variables, and let $f : \mathbb{C}^n \to \mathbb{R}$ be a 1-Lipschitz function. Then for any t one has

$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{E}f(\mathbf{X})| \geqslant tK\right) \leqslant C\exp\left(-ct^2\right)$$

for some absolute constants $C, c > 0$.

The theorem is valid for all Lipschitz functions for Gaussian random vectors.

Theorem 1.3.5 (Talagrand concentration inequality) Let $K > 0$, and let $X_1, \ldots, X_n$ be independent complex variables with $|X_i| \leqslant K$ for all $1 \leq i \leq n$. Let $f : \mathbb{C}^n \to \mathbb{R}$ be a 1-Lipschitz and convex function. Then for any t one has

$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{M}f(\mathbf{X})| \geqslant tK\right) \leqslant C\exp\left(-ct^2\right)$$
$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{E}f(\mathbf{X})| \geqslant tK\right) \leqslant C\exp\left(-ct^2\right)$$

for some absolute constants $C, c > 0$, where $\mathbb{M}f(\mathbf{X})$ is a median of $f(\mathbf{X})$

The theorem is valid for all Lipschitz and convex functions for independent (not necessarily Gaussian) random vectors.

Example 1.3.6 (large random matrix theory) Random matrix theory or quantum information theory is very relevant for big data. The vision of exploiting random matrixes to model big data is explicitly proposed in [39].

For N random (row) vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{C}^{1\times T}$, we form an $N \times T$ random matrix

$$\mathbf{X} = \left(\mathbf{X}_1, \ldots, \mathbf{X}_N\right)^T \in \mathbb{C}^{N\times T}$$

We say a matrix $\mathbf{Y}$ is Hermitian if $\mathbf{Y} = \mathbf{Y}^H$, where H denotes the conjugate and transpose of a matrix. In general, the random matrix $\mathbf{X}$ is not Hermitian.

The classical framework is to study the regime of N fixed while $T \to \infty$. For modern big data, this fundamental assumption is invalid. We must study the new paradigm

$$N \to \infty, T \to \infty \text{ but N/T} \to c \in [0, \infty)$$

where c is a fixed constant.

This book surveys a lot of recent results of the literature in Part I. □

Example 1.3.7 (free probability theory for hermitian random matrices) When the sizes of random matrices are large, conventional independence is replaced with asymptotically freeness. Free random variables may be thought of as "independent" random matrices in the classical sense. Chapter 5 applies this theory to model large random matrices. Free random variables are random infinite-dimensional linear operators that are equivalently very large random matrices. The statistical properties of free random variables are equivalently those of the eigenvalues of large random matrices. □

Free probability theory was introduced by Voiculescu around 1983 in order to attack the isomorphism problem of von Neumann algebras of free groups. Voiculescu isolated a structure showing up in this context, which he named "freeness." His fundamental insight was to separate this concept from its operator algebraic origin and investigate it for its own sake. Furthermore, he promoted the point of view that freeness should be seen as a noncommutative analog of the classical probabilistic concept of "independence" for random variables. Hence freeness is also called "free independence" and the whole subject became known as "free probability theory."

The theory was lifted to a new level when Voiculescu discovered, in 1991, that the freeness property is also present for many classes of *random matrices*, in the asymptotic regime when the size of the matrices tends to infinity. This insight, bringing together the a priori entirely different theories of operator algebras and of random matrices, had quite some impact in both directions. Modeling operator algebras by random matrices resulted in some insightful results about operator algebras, whereas tools developed in operator algebras and free probability theory could now be applied to random matrix problems, yielding, in particular, new ways to calculate the asymptotic eigenvalue distribution of many random matrices. Freeness is motivated not by its initial occurrence in operator algebras but by its random matrix connection.

In free probability theory, the central limit theorem on the sum of independent free random variables gives a semicircle distribution. A semicircle distribution serves the same function as the Gaussian or normal distribution for the sum of independent commuting random variables. If $X_1, X_2, \ldots, X_n$ are identically distributed zero mean free random variables with variance of $(R/2)^2$, the free summation or additive free convolution of

$$\frac{1}{\sqrt{n}} X_1 \boxplus X_2 \boxplus \cdots \boxplus X_n$$

has the semicircle distribution of

$$p(t) = \begin{cases} \frac{1}{2\pi R^2}\sqrt{R^2 - t^2} & |t| \leqslant R \\ 0 & \text{otherwise,} \end{cases}$$

where R is the radius of the distribution and $\boxplus$ denotes the additive free convolution.

Using free probability, we can calculate the histogram for a generic realization of a 3000×3000 random matrix $p(X, Y)$, where X and Y are, respectively, independent Gaussian and Wishart random matrices: $p(X, Y) = X + Y$; $P(X, Y) = XY + YX + X^2$. $P(X, Y)$ is a polynomial of two random matrices.

Example 1.3.8 (free probability theory for non-Hermitian random matrices) As pointed out above, in general, a random matrix $\mathbf{Y}$ is non-Hermitian. Most tools in algebra deal with Hermitian random matrices. Non-Hermitian random matrices are much more difficult to handle, compared with their Hermitian counterparts. Chapter 6 gives a comprehensive introduction to model the data using (large) non-Hermitian random matrices. □

The eigenvalue density of a product

$$\mathbf{X}_1\mathbf{X}_2\cdots\mathbf{X}_L \tag{1.3}$$

of $L \geq 2$ independent $N \times N$ Gaussian random matrices in the limit $N \to \infty$ is rotationally symmetric in the complex plane and is given by a simple expression

$$\rho\left(z,\overline{z}\right) = \begin{cases} \frac{1}{L\pi}\sigma^{-2/L}|z|^{-2+2/L} & |z| \leqslant \sigma \\ 0 & |z| > \sigma \end{cases}$$

where the $\overline{z}$ denotes the complex conjugate of a complex number z, and the effective scale parameter $\sigma = \sigma_1\sigma_2\cdots\sigma_L$. We have

$$\mathbb{E}\left(\mathbf{X}_1\right)_{ij} = \cdots = \mathbb{E}\left(\mathbf{X}_L\right)_{ij} = 0, \quad i,j = 1,\ldots,N$$
$$\mathbb{E}\left|\left(\mathbf{X}_1\right)_{ij}\right|^2 = \sigma_1^2/N, \cdots, \mathbb{E}\left|\left(\mathbf{X}_L\right)_{ij}\right|^2 = \sigma_L^2/N, \quad i,j = 1,\ldots,N$$

The parameter σ corresponds to the radius of the circular support and is related to the amplitude of the Gaussian fluctuations. This form of the eigenvalue density is universal. It is identical for products of Gaussian Hermitian, non-Hermitian, real or complex random matrices. It does not change even if the matrices in the product are taken from different Gaussian ensembles.

Study the product

$$\mathbf{P} = \mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_L \tag{1.4}$$

of $L \geq 1$ independent rectangular large random Gaussian matrices $\mathbf{A}_l, l = 1,2,\ldots,L$ of dimensions $N_l \times N_{l+1}$. We are interested in the limit $N_{L+1} \to \infty$ and

$$R_l \equiv \frac{N_l}{N_{l+1}} = \text{finite}, \quad \text{for } l = 1,2,\ldots,L+1$$

The σ_l parameters set the scale for the Gaussian fluctuations in $\mathbf{A}_l$s. The entries of each matrix $\mathbf{A}_l$ can be viewed as independent centered Gaussian random variables, the variance of the real and imaginary parts being proportional to σ_l^2 and inversely proportional to the square root of the number N_lN_{l+1} of elements in the matrix.

Consider

$$\mathbf{Q} = \mathbf{P}^H\mathbf{P}, \qquad \mathbf{R} = \mathbf{P}\mathbf{P}^H$$

where $\mathbf{P}$ is defined in (1.4). $\mathbf{Q}$ and $\mathbf{R}$ are are Hermitian, and they have non-negative spectra, which differ only in the zero modes. The M transform of the matrix $\mathbf{X}$ is defined as

$$M_{\mathbf{X}}\left(z,\overline{z}\right)=zG_{\mathbf{X}}\left(z,\overline{z}\right)-1$$

where $G_{\mathbf{X}}\left(z,\overline{z}\right)$ is the Green's function.

The main finding is that the eigenvalue distribution and the M transform of the product [(1.4)] are spherically symmetric. We shall show the M transform to satisfy the L-th order polynomial equation:

$$\prod_{l=1}^{L}\left(\frac{M_{\mathbf{P}}\left(|z|^2\right)}{R_l}+1\right)=\frac{|z|^2}{\sigma^2} \tag{1.5}$$

where the scale parameter is $\sigma=\sigma_1\sigma_2\cdots\sigma_M$.

An analogous equation for $\mathbf{Q}$ reads

$$\sqrt{R_l}\frac{M_{\mathbf{Q}}(z)+1}{M_{\mathbf{Q}}(z)}\prod_{l=1}^{L}\left(\frac{M_{\mathbf{Q}}(z)}{R_l}+1\right)=\frac{z}{\sigma^2} \tag{1.6}$$

The free argument in (1.5) is $|z|^2$, and z in (1.6). It is surprising that there is rotational symmetry in the complex plane for the product of Gaussian random matrices $\mathbf{P}$, while the study of the Hermitian product $\mathbf{Q}$ breaks the rotational symmetry. In other words, given a data matrix $\mathbf{A}_l, l=1,\ldots,L$, some statistical structure (symmetry) will be lost if we study the non-negative Hermitian random matrix $\mathbf{Q}$, instead of non-Hermitian random matrix $\mathbf{P}$.

One unexpected implication of the universality is that a product of random matrices whose spectra do not necessarily display rotational symmetry has an eigenvalue distribution that does possess rotational symmetry on the complex plane (i.e., the average density depends only on $|\lambda|$).

A random quantum state is defined by specifying a probability measure in the space of density matrices $\boldsymbol{\rho}$, i.e., Hermitian, weakly positive-definite (i.e., with nonnegative eigenvalues), and normalized (i.e., $\mathrm{Tr}\,\boldsymbol{\rho}=1$) matrices. For any rectangular matrix $\mathbf{Z}$, one can define $\boldsymbol{\rho}\equiv\mathbf{Z}\mathbf{Z}^H/\mathrm{Tr}\left(\mathbf{Z}\mathbf{Z}^H\right)$ is a proper random quantum density matrix.

If we model the system using the random states through $\boldsymbol{\rho}$, we will break the rotational symmetry of the product of random matrices $\mathbf{P}$ (defined in (1.4)) in the complex plane. It makes sense because the eigenvalues of $\mathbf{P}$ are distributed in the complex plane and the eigenvalues of $\boldsymbol{\rho}$ are in the real axis (non-negative real values).

In statistics we often use a sample covariance matrix in the form of $\mathbf{Q}$. The comments for $\boldsymbol{\rho}$ are also valid for a sample covariance matrix. By studying the sample covariance matrix we will lose some structure information (such as rotational symmetry in the complex plane for the M transform). See Example 1.3.9 for the potential relevance to applications.

We now consider the eigenvalue statistics for complex $N\times N$ Wishart matrices $\mathbf{X}_{r,s}^H\mathbf{X}_{r,s}$, where $\mathbf{X}_{r,s}$ is equal to the product of r complex Gaussian matrices, and the inverse of s complex Gaussian matrices. In particular, we have

$$\mathbf{X}_{r,s}=\mathbf{G}_r\mathbf{G}_{r-1}\cdots\mathbf{G}_1\left(\tilde{\mathbf{G}}_s\tilde{\mathbf{G}}_{s-1}\cdots\tilde{\mathbf{G}}_1\right)^{-1}$$

where each $\mathbf{G}_k$ is a rectangular standard complex Gaussian matrix of dimension $n_k \times n_{k-1}, n_k \geqslant n_{k-1},$ and $n_0 = N$, and each $\tilde{\mathbf{G}}_k$ is a square of dimension $N \times N$.

Example 1.3.9 (functional averages over Gaussian ensembles) The MIMO channel model is defined similarly to (3.11). The result here can be applied to massive MIMO analysis. See Section 15.3. We repeat the definition to fix a different notation. Denoting the number of transmitting antennas by M and the number of receiving antennas by N, the channel model is

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n} \tag{1.7}$$

where $\mathbf{s} \in \mathbb{C}^M$ is the transmitted vector, $\mathbf{y} \in \mathbb{C}$ is the received vector, $\mathbf{H} \in \mathbb{C}^{N \times M}$ is a complex matrix and $\mathbf{n} \in \mathbb{C}^N$ is the zero mean complex Gaussian vector with independent, equal variance entries. We assume that $\mathbb{E}\left(\mathbf{n}\mathbf{n}^H\right) = \mathbf{I}_N$, where $(\cdot)^H$ denotes the complex conjugate transpose and $\mathbf{I}_N$ the $N \times N$ identity matrix. It is reasonable to put a power constraint

$$\mathbb{E}\left(\mathbf{n}^H\mathbf{n}\right) = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{n}\mathbf{n}^H\right)\right] \leqslant P$$

where P is the total transmitted power. The signal-to-noise ratio, denoted by snr, is defined as the quotient of the signal power and the noise power, and in this case is equal to P/N. □

Recall that if $\mathbf{A}$ is an $n \times n$ Hermitian matrix then there exists $\mathbf{U}$ unitary and $\mathbf{D} = \mathrm{diag}\left(d_1, \ldots, d_n\right)$ such that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^H$. Given a continuous function f, we define $f(\mathbf{A})$ as

$$f(\mathbf{A}) = \mathbf{U}\,\mathrm{diag}\left(f\left(d_1\right), \ldots, f\left(d_n\right)\right)\mathbf{U}^H$$

Naturally, the simplest example is the one where $\mathbf{H}$ has independent and identically distributed (i.i.d.) Gaussian entries, which constitutes the canonical model for the single-user narrow band MIMO channel. It is known that the capacity of this channel is achieved when $\mathbf{s}$ is a vector with complex Gaussian zero mean and covariance $\mathrm{snr}\,\mathbf{I}_M$. See [51, 52] for instance. For the fast fading channel, assuming statistical channel state information at the transmitter, the ergodic capacity is given by

$$\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] = \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \tag{1.8}$$

where in the last equality we use the fundamental fact that

$$\log\det(\cdot) = \mathrm{Tr}\log(\cdot) \tag{1.9}$$

We prefer the form of $\mathrm{Tr}\log(\cdot)$ because the trace $\mathrm{Tr}(\cdot)$ is a linear function. The expectation $\mathbb{E}(\cdot)$ is also a linear function. Sometimes it is convenient to exchange the order of $\mathbb{E}$ and $\mathrm{Tr}(\cdot)$ in (1.8):

$$\begin{aligned}\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] &= \mathbb{E}\left[\mathrm{Tr}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &= \mathrm{Tr}\left[\mathbb{E}\log\left(\mathbf{I}_N + \mathrm{snr}\,\mathbf{H}\mathbf{H}^H\right)\right]\end{aligned}$$

The $\mathbb{E}(\mathbf{X})$ can be approximated by the arithmetic average $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$ when n "snapshots" of the $p \times p$ random matrix $\mathbf{X}$ are observed. As a result, we reach

$$\begin{aligned}\mathbb{E}\left[\log\det\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] &= \mathbb{E}\left[\text{Tr}\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &= \text{Tr}\left[\mathbb{E}\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}\mathbf{H}^H\right)\right] \\ &\approx \frac{1}{n}\,\text{Tr}\left[\sum_{i=1}^{n}\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}_i\mathbf{H}_i^H\right)\right] \end{aligned} \tag{1.10}$$

which boils down to the sum of random positive definite Hermitian matrices $\mathbf{H}_i\mathbf{H}_i^H$, $i = 1, \ldots, n$, given the i-th "snapshot" $\mathbf{H}_i$ of the random channel matrix $\mathbf{H}$ that is defined in (3.16). See [40] for a whole chapter on the sum of random matrices. The channel capacity with a finite number of samples can be obtained using (1.10). Note that the Frobenius norm is defined as

$$\|\mathbf{B}\|_F^2 \equiv \text{Tr}\left(\mathbf{B}\mathbf{B}^H\right)$$

In (1.10), if we expand the function $\log\left(\mathbf{I}_N + \text{snr}\,\mathbf{H}_i\mathbf{H}_i^H\right)$ using its Taylor series, we can reduce the problem to the sample moments m_k defined as

$$\hat{m}_k = \frac{1}{M}\,\text{Tr}\left[\left(\frac{1}{N}\mathbf{H}_i\mathbf{H}_i^H\right)^k\right]$$

for an integer $k \geq 1$. Because the sample moments $\hat{m}_k$ are *consistent estimators* of true moments m_k, it is then natural to use the moment method for the inference of the parameters [53, p. 425]. See Section 8.9.3 for this connection.

More generally, we can expand a functional of a random matrix in the form of $f\left(\mathbf{H}\mathbf{H}^H\right)$ in terms of its Taylor series. We can similarly obtain the true moments m_k. We can use sample moments $\hat{m}_k$ to estimate the true moments.

Another important performance measure is the minimum mean square error (MMSE) achieved by a linear receiver, which determines the maximum achievable output signal to interference and noise ratio (SINR). For an input vector $\mathbf{x}$ with i.i.d. entries of zero mean and unit variance, the MSE at the output of the MMSE receiver is given by

$$\min_{\mathbf{M}\in\mathbb{C}^{M\times N}}\mathbb{E}\left[\|\mathbf{x} - \mathbf{M}\mathbf{y}\|^2\right] = \mathbb{E}\left[\text{Tr}\log\left(\mathbf{I}_M + \text{snr}\,\mathbf{H}^H\mathbf{H}\right)^{-1}\right] \tag{1.11}$$

where the expectation on the left-hand side is over both the vectors x and the random matrices $\mathbf{H}$, whereas the right-hand side is over $\mathbf{H}$ only. See [52] for details.

Let $\mathbf{H}$ be an $n \times n$ Gaussian random matrix with complex, independent, and identically distributed entries of zero mean and unit variance. Given an $n \times n$ positive definite matrix $\mathbf{A}$, and a continuous function $f : \mathbb{R}^+ \to \mathbb{R}$ such that $\int_0^\infty e^{-\alpha t}|f(t)|^2 dt < \infty$ for every $\alpha > 0$, Tucci and Vega (2013) [54] find a new formula for the expectation

$$\mathbb{E}\left[\text{Tr}\left(f\left(\mathbf{H}\mathbf{A}\mathbf{H}^H\right)\right)\right]$$

Taking $f(x) = \log(1 + x)$ gives another formula for the capacity of the MIMO communication channel, and taking $f(x) = (1 + x)^{-1}$ gives the MMSE achieved by a linear receiver.

From Example 1.3.8, we see the connection of eigenvalues of $\mathbf{H}$ and $\mathbf{H}\mathbf{H}^H$, when $\mathbf{H}$ is decomposed into a product of L random matrices.

Example 1.3.10 (matrix hypothesis testing) Applications include: (i) anomaly detection; (ii) denial of service for big data; (iii) bad data detection for Smart Grid (state estimation). We consider the so-called matrix hypothesis-testing problem

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{Y} = \mathbf{X} \\ \mathcal{H}_1 &: \quad \mathbf{Y} = \sqrt{\text{SNR}} \cdot \mathbf{H} + \mathbf{X} \end{aligned} \tag{1.12}$$

where SNR represents the signal-to-noise ratio, and $\mathbf{X}$ is a non-Hermitian random matrix of $m \times n$. We further assume that $\mathbf{H}$ is independent of $\mathbf{X}$. The problem of (1.12) is equivalent to

$$\begin{aligned} \mathcal{H}_0 &: \quad \mathbf{Y}\mathbf{Y}^H = \mathbf{X}\mathbf{X}^H \\ \mathcal{H}_1 &: \quad \mathbf{Y}\mathbf{Y}^H = \text{SNR} \cdot \mathbf{H}\mathbf{H}^H + \mathbf{X}\mathbf{X}^H + \sqrt{\text{SNR}}\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right) \end{aligned} \tag{1.13}$$

where $\mathbf{H}\mathbf{H}^H, \mathbf{X}\mathbf{X}^H, \mathbf{Y}\mathbf{Y}^H$ are positive semidefinite Hermitian random matrices, which are Wishart matrices if $\mathbf{X}, H$ are Gaussian random matrices. A matrix $\mathbf{A}$ of $m \times n$ is said to be positive semidefinite if all the eigenvalues of $\mathbf{A}$ are non-negative, i.e., $\lambda_i(\mathbf{A}) \geq 0, i = 1, \ldots, \min(m, n)$. The matrix $\left(\mathbf{H}\mathbf{X}^H + \mathbf{X}\mathbf{H}^H\right)$ is Hermitian. □

The likelihood ratio test (LRT) is the natural choice. We deal with matrix-valued random variables, where the matrix sizes are large. See Section 8.11 for details. The analysis of these metrics requires advanced tools, such as the nonasymptotic theory of random matrices. The nonasymptotic theory is based on the "concentration of measure" phenomenon when the size of a matrix is large but finite. This phenomenon is the starting point for almost all the results.

Theorem 17.3.1 essentially says that if we take two large random matrices $\mathbf{A}_N$ and $\mathbf{B}_N$, and if we conjugate one of them by a uniformly random unitary transformation $\mathbf{U}_N$, then the resulting pair of matrices $\mathbf{A}_N$ and $\mathbf{U}_N\mathbf{B}_N\mathbf{U}_N^H$ will be approximately free. As a slogan, this can be expressed as follows

Two large random matrices
in the general position
are asymptotically free!

For a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is well known that the differential entropy $H(\cdot)$ is given by

$$\mathcal{H}(\boldsymbol{\Sigma}) = \frac{p}{2} + \frac{1}{2}p\log(2\pi) + \frac{1}{2}\log\det\boldsymbol{\Sigma}. \tag{1.14}$$

The high-dimensional setting where the dimension $p(n)$ grows with the sample size n is of particular current interest.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_{n+1}$ be an independent random sample from the p-dimensional Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The sample covariance matrix is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n+1}\left(\mathbf{X}_k - \overline{\mathbf{X}}\right)\left(\mathbf{X}_k - \overline{\mathbf{X}}\right)^T$$

A central limit theorem is established for the log determinant of $\hat{\mathbf{\Sigma}}$ in the high-dimensional setting where the dimension p grows with the sample size n with the only restriction that $p(n) \leq n$. In the case when $\lim_{n\to 0} \frac{p(n)}{n} = r$ for some $0 \leq r \leq 1$, the central limit theorem shows

$$\frac{\log\det\hat{\mathbf{\Sigma}} - \sum_{k=1}^{p} \log\left(1 - \frac{k}{n}\right) - \log\det\mathbf{\Sigma}}{\sqrt{-2\log\left(1 - \frac{p}{n}\right)}} \xrightarrow{Law} \mathcal{N}(0,1) \text{ as } n \to \infty \tag{1.15}$$

The result for the boundary case $p = n$ yields

$$\frac{\log\det\hat{\mathbf{\Sigma}} - \log(n-1)! + n\log n - \log\det\mathbf{\Sigma}}{\sqrt{2\log n}} \xrightarrow{Law} \mathcal{N}(0,1) \text{ as } n \to \infty \tag{1.16}$$

One common problem in statistics and engineering is to estimate the distance between two population distributions based on the samples. A commonly used measure of closeness is the relative entropy or the Kullback–Leibler divergence. For two distributions $\mathbb{P}$ and $\mathbb{Q}$ with respective density functions $p(\cdot)$ and $q(\cdot)$, the relative entropy between $\mathbb{P}$ and $\mathbb{Q}$ is

$$KL(\mathbb{P},\mathbb{Q}) = \int p(x)\log\frac{p(x)}{q(x)}dx$$

In the case of two multivariate Gaussian distributions $\mathbb{P} = \mathcal{N}_p\left(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\right)$, $\mathbb{Q} = \mathcal{N}_p\left(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2\right)$

$$2KL(\mathbb{P},\mathbb{Q}) = \mathrm{Tr}\left(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1\right) - p + \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^T \mathbf{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right) + \log\left(\frac{\det\mathbf{\Sigma}_1}{\det\mathbf{\Sigma}_2}\right) \tag{1.17}$$

From (1.17), it is clear that estimation of the relative entropy involves estimation of the log determinants $\log\det\mathbf{\Sigma}_1$ and $\log\det\mathbf{\Sigma}_2$.

For testing the hypothesis that two multivariate Gaussian distributions $\mathbb{P} = \mathcal{N}_p\left(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\right)$, and $\mathbb{Q} = \mathcal{N}_p\left(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2\right)$ have the same entropy, we have

$$\mathcal{H}_0 : \mathcal{H}(\mathbb{P}) = \mathcal{H}(\mathbb{Q}) \text{ versus } \mathcal{H}_1 : \mathcal{H}(\mathbb{P}) \neq \mathcal{H}(\mathbb{Q})$$

For any given significance level $0 < \alpha < 1$, a test with the asymptotic level α can be constructed easily using the central limit theorem given above, based on two independent samples, one from $\mathbb{P}$ and another from $\mathbb{Q}$.

Knowledge of the log determinant of covariance matrices is also essential for the quadratic discriminant analysis (QDA). For classification of two multivariate Gaussian distributions $\mathcal{N}_p\left(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\right)$ and $\mathcal{N}_p\left(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2\right)$, when the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ are known, the oracle discriminant is

$$\Delta = -\left(\mathbf{z} - \boldsymbol{\mu}_1\right)^T \mathbf{\Sigma}_1^{-1}\left(\mathbf{z} - \boldsymbol{\mu}_1\right) + \left(\mathbf{z} - \boldsymbol{\mu}_2\right)^T \mathbf{\Sigma}_2^{-1}\left(\mathbf{z} - \boldsymbol{\mu}_2\right) - \log\left(\frac{\det\mathbf{\Sigma}_1}{\det\mathbf{\Sigma}_2}\right) \tag{1.18}$$

That is, the observation vector z is classified into the population with $\mathcal{N}_p\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ distribution if $\Delta > 0$ and into $\mathcal{N}_p\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$ otherwise.

Example 1.3.11 (outliers in signal plus noise) For a complex variable $z = x + iy$, the Dirac delta function is defined by $\delta^2(z) \equiv \delta(x)\,\delta\left(y\right)$, and we define $\partial/\partial\overline{z} = \left(\partial/\partial x + i\partial/\partial y\right)/2$, and $\partial/\partial z = \left(\partial/\partial x - i\partial/\partial y\right)/2$. For simplicity, we use the notation $f(z)$ (instead of $f\left(z, \overline{z}\right)$) for general, nonholomorphic functions on the complex plane. □

We provide a general formula for the eigenvalue density of large random $N \times N$ matrices of the form

$$\mathbf{A} = \mathbf{M} + \mathbf{LXR} \tag{1.19}$$

where $\mathbf{M}$, $\mathbf{L}$ and $\mathbf{R}$ are general ($\mathbf{M}$) or arbitrary invertible ($\mathbf{L}$ and $\mathbf{R}$) deterministic matrices, and $\mathbf{X}$ is a random matrix of zero-mean independent and identically distributed (i.i.d.) elements with zero mean and variance $1/N$. For example, the entries of $\mathbf{X}$ are Gaussian or Bernoulli random variables. The model (1.19) has been used to model the brain, and may be used for sensor networks and wireless networks.

As $\mathbf{X}$ and therefore $\mathbf{L}XR$ have zero mean, $\mathbf{M}$ is the ensemble average of $\mathbf{A}$. The random fluctuations of $\mathbf{A}$ around its average are given by the matrix $\mathbf{L}XR$, which for general $\mathbf{L}$ and/or $\mathbf{R}$ has dependent and nonidentically distributed elements, due to the possible mixing and nonuniform scaling of the rows (columns) of the i.i.d. $\mathbf{X}$ by $\mathbf{L}$ ($\mathbf{R}$).

The density of the eigenvalues of $\mathbf{A} = \mathbf{M} + \mathbf{LXR}$, in the complex plane for a realization of $\mathbf{X}$ (also known as the empirical spectral distribution), is defined by

$$\rho_{\mathbf{X}}(z) = \frac{1}{N}\sum_{i=1}^{N}\delta^2\left(z - \lambda_i\right)$$

where λ_i are the eigenvalues of $\mathbf{M} + \mathbf{LXR}$. It is known [55] that $\rho_{\mathbf{X}}(z)$ is asymptotically self-averaging, in the sense that with probability 1, $\rho_{\mathbf{X}}(z) - \rho(z)$ converges to zero (in the distributional sense) as $N \to \infty$, where $\rho(z) \equiv \langle \rho_{\mathbf{X}}(z) \rangle_{\mathbf{X}}$ is the ensemble average of $\rho_{\mathbf{X}}(z)$. Thus for large enough N, any typical realization of $\mathbf{X}$ yields an eigenvalue density $\rho_{\mathbf{X}}(z)$ that is arbitrarily close to $\rho(z)$.

For any matrix $\mathbf{B}$, we denote its operator norm (its maximum singular value) by $\|\mathbf{B}\|$ and we define its (normalized) Frobenius norm via

$$\|\mathbf{B}\|_F \equiv \frac{1}{N}\sum_{i,j=1}^{N}\left|B_{ij}\right|^2 = \frac{1}{N}\,\mathrm{Tr}\left(\mathbf{B}\mathbf{B}^H\right) \tag{1.20}$$

(equivalently, $\|\mathbf{B}\|_F$ is the root mean square of the singular values of $\mathbf{B}$).

Our general result is that for large N, $\rho(z)$ is nonzero in the region of complex plane satisfying

$$\frac{1}{N}\,\mathrm{Tr}\left[\left(\mathbf{M}_z\mathbf{M}_z^{\dagger}\right)^{-1}\right] \geqslant 1 \tag{1.21}$$

where we defined

$$\mathbf{M}_z = L^{-1}\left(z\mathbf{I} - \mathbf{M}\right)\mathbf{R}^{-1}$$

Using (1.20) , we can express (1.21) as

$$\left\|\mathbf{R}(z\mathbf{I}-\mathbf{M})^{-1}\mathbf{L}\right\|_F \geqslant 1$$

inside this region, $\rho(z)$ is given by

$$\rho(z) = \frac{1}{N}\frac{1}{z}\frac{\partial}{\partial \overline{z}} \operatorname{Tr}\left[(\mathbf{RL})^{-1}\mathbf{M}_z^H\left(\mathbf{M}_z\mathbf{M}_z^H + g(z)^2\right)^{-1}\right] \tag{1.22}$$

where $g(z)$ is a real, scalar function found by solving

$$\frac{1}{N}\operatorname{Tr}\left[\left(\mathbf{M}_z\mathbf{M}_z^H + g^2\right)^{-1}\right] = 1,$$

for g for each z.

Example 1.3.12 (asymptotically deterministic character of limiting spectral distributions) One motivation is to study the random block matrices. We consider $N \times N$ matrices that are Hermitian with above diagonal "block-rows" (or "strips") of height bounded by a constant d. Examples are matrices with i.i.d block entries but the theory we develop here applies more generally. □

Our analysis is often based on Stieltjes transforms. We call

$$m_n(z) = \frac{1}{N}\operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

the Stieltjes transform of $\mathbf{M}$, the $N \times N$ random matrix of interest. Here $\mathbf{I}_N$ is the $N \times N$ identity matrix. In much of our analysis, we will let N grow to infinity.

Theorem 1.3.13 Suppose the $N \times N$ Hermitian matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \sum_{i=1}^{n} \mathbf{M}_i,$$

where $\mathbf{M}_i$ are *independent* with $\operatorname{rank}\left(\mathbf{M}_i\right) \leqslant d_i$. Let $z \in \mathbb{C}^+$ and $\operatorname{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N}\operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left|m_n(z) - \mathbb{E}\left(m_n(z)\right)\right| > t\right) \leqslant C\exp\left(-c\frac{N^2v^2t^2}{\sum_{i=1}^{n} d_i^2}\right)$$

where C and c are two constants that do not depend on n or d_is.

We can extend the above theorem to the following.

Theorem 1.3.14 Suppose the $N \times N$ Hermitian matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \sum_{1 \leqslant i,j \leqslant n} \mathbf{\Theta}_{i,j}$$

where $\mathbf{\Theta}_{i,j} = f_{i,j}\left(Z_i, Z_j\right)$ is a $N \times N$ matrix and the random variables $\left\{Z_i\right\}_{i=1}^{n}$ are independent. ($f_{i,j}\left(Z_i, Z_j\right)$ are simply matrix valued functions of our random variables.) Let $\mathbf{M}_i$ be the Hermitian matrix

$$\mathbf{M}_i = \mathbf{\Theta}_{i,i} + \sum_{j \neq i} \left(\mathbf{\Theta}_{i,j} + \mathbf{\Theta}_{j,i}\right)$$

Assume that $\operatorname{rank}\left(\mathbf{M}_i\right) \leqslant d_i$. Let $z \in \mathbb{C}^+$ and $\operatorname{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N} \operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left|m_n(z) - \mathbb{E}\left(m_n(z)\right)\right| > t\right) \leqslant C \exp\left(-c \frac{N^2 v^2 t^2}{\sum_{i=1}^{n} d_i^2}\right)$$

where C and c are two constants that do not depend on n nor d_is.

The previous theorem is derived from the following theorem.

Theorem 1.3.15 Suppose that the $N \times N$ Hermitian matrix $\mathbf{M}$ is such that, for independent random variables $\left\{Z_i\right\}_{i=1}^{n}$ and a matrix valued function f,

$$\mathbf{M} = f\left(Z_1, \ldots, Z_n\right)$$

Suppose further that for all $1 \leq i \leq n$, there exists a matrix $\mathbf{N}_i$ such that

$$\mathbf{N}_i = f_i\left(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\right)$$

and $\operatorname{rank}\left(\mathbf{M} - \mathbf{N}_i\right) \leqslant d_i$. (When $i = 1$, $\mathbf{N}_1 = f_1\left(Z_2, \ldots, Z_n\right)$ and when $i = n$, $\mathbf{N}_n = f_n\left(Z_1, \ldots, Z_{n-1}\right)$. f_is are simply matrix-valued functions.) Let $z \in \mathbb{C}^+$ and $\operatorname{Im}[z] = v > 0$. Call

$$m_n(z) = \frac{1}{N} \operatorname{Tr}\left(\left(\mathbf{M} - z\mathbf{I}_N\right)^{-1}\right)$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\left|m_n(z) - \mathbb{E}\left(m_n(z)\right)\right| > t\right) \leqslant C \exp\left(-c \frac{N^2 v^2 t^2}{\sum_{i=1}^{n} d_i^2}\right)$$

where C and c are two constants that do not depend on n nor d_is.

This is McDiarmid-type inequality.

Example 1.3.16 (random particles) Beyond random matrices, how about the empirical measure of random particles in $\mathbb{R}^d$? Is there an analog of the circular law phenomenon? Does the ball replace the disc? The answer is positive. A wireless radio sensor can be modeled as a random particle, for example. □

We consider a system of N particles in $\mathbb{R}^d$ at positions $\mathbf{x}_1, \ldots, \mathbf{x}_N$, say with charge $1/N$. These particles are subject to confinement by an external field via a potential $\mathbf{x} \in \mathbb{R}^d \mapsto V(\mathbf{x})$. and to internal pair interaction (typically repulsion) via a potential $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto W(\mathbf{x}, \mathbf{y})$. The idea is that an equilibrium may emerge as N tends to infinity. The configuration energy is

$$\begin{aligned}\mathcal{I}_N(\mathbf{x}_1, \ldots, \mathbf{x}_N) &= \frac{1}{N}\sum_{i=1}^{N} V(\mathbf{x}_i) + \frac{1}{N^2}\sum_{1\leqslant i<j\leqslant N} W(\mathbf{x}_i, \mathbf{x}_j) \\ &= \int V(\mathbf{x})\, d\mu_N(\mathbf{x}) + \frac{1}{2}\int_{\neq} W(\mathbf{x}, \mathbf{y})\, d\mu_N(\mathbf{x})\, d\mu_N(\mathbf{y})\end{aligned}$$

where μ_N is the empirical measure of the particles (global encoding of the particle system)

$$\mu_N := \frac{1}{N}\sum_{k=1}^{N} \delta_{\mathbf{x}_k}$$

The model is mean field in the sense that each particle interacts with the others only via the empirical measure of the system. If $1 \leq d \leq 2$, then one can construct a random normal matrix which admits our particles at $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as eigenvalues: for any $n \times n$ unitary matrix $\mathbf{U}$,

$$\mathbf{M} = \mathbf{U}\,\mathrm{diag}(\mathbf{x}_1, \ldots, \mathbf{x}_N)\,\mathbf{U}^H$$

which is unitary invariant if $\mathbf{U}$ is Haar distributed. Here we are more interested in an arbitrarily high dimension d, for which no matrix model is available. We make our particles at $\mathbf{x}_1, \ldots, \mathbf{x}_N$, random by considering the exchangeable probability measure P_N on $(\mathbb{R}^d)^N$ with density proportional to

$$\exp\left(-\beta_N \mathcal{I}_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)\right)$$

where $\beta_N > 0$ is a positive parameter that may depend on N. The law P_N is a Boltzmann measure at inverse temperature β_N, and takes the form $\prod_{i=1}^{N} f_1(\mathbf{x}_i) \prod_{1\leqslant i<j\leqslant N} f_2(\mathbf{x}_i, \mathbf{x}_j)$ due to the structure and symmetries of $\mathcal{I}_N$.

The model contains the complex Ginibre ensemble of random matrices as the special case

$$d = 2, \beta_N = N^2, V(\mathbf{x}) = |\mathbf{x}|^2, W(\mathbf{x}, \mathbf{y}) = 2\log\frac{1}{|\mathbf{x} - \mathbf{y}|}$$

which is two dimensional, with quadratic confinement, Coulomb repulsion, and temperature $1/N^2$. Here we denote by $|\cdot|$ the Euclidean norm of $\mathbb{R}^d$.

Beyond this two-dimensional example, the typical interaction potential W that we may consider is the Coulomb interaction in arbitrary dimension

$$W(\mathbf{x},\mathbf{y}) = K_{\Delta}(\mathbf{x}-\mathbf{y}) \text{ with } K_{\Delta}(\mathbf{x}) = \begin{cases} |\mathbf{x}| & ifd = 1 \\ \log\frac{1}{|\mathbf{x}|} & ifd = 2 \\ \frac{1}{|\mathbf{x}|^{d-2}} & ifd \geqslant 3 \end{cases}$$

and the Riesz interaction, $0 < \alpha < d$ (Coulomb if $d \geq 3$ and $\alpha = 2$) $d \geq 1$

$$W(\mathbf{x},\mathbf{y}) = K_{\Delta_\alpha}(\mathbf{x}-\mathbf{y}) \text{ with } K_{\Delta_\alpha}(\mathbf{x}) = \frac{1}{|\mathbf{x}|^{d-\alpha}}$$

The Coulomb kernel K_{Δ} is the fundamental solution of the Laplace equation, whereas the Riesz kernel K_{Δ_α} is the fundamental solution of the fractional Laplace equation, hence the notations. In other words, in the sense of Schwartz–Sobolev distributions, for some constant c_d,

$$\Delta_\alpha K_{\Delta_\alpha} = c_d \delta_0$$

If $\alpha \neq 2$, then the operator Δ_α is a nonlocal Fourier multiplier.

1.4 A Mathematical Theory of Big Data

This section presents a mathematical theory to unify big data systems. Basic questions for big data includes:

- What is the theoretical foundation of big data?
- The science of data or the science of information?
- What is information?
- Are the definitions of information given by Shannon and Von Neumann sufficient for big data?
- How is "free entropy" relevant to the new definition of "information"?

Applications of big data include: (i) quantum systems; (ii) financial systems; (iii) atmospheric systems; (iv) sensor network (e.g., PMU, WAMS); (v) wireless networks (vehicle-to-vehicle communications, 5G); (vi) transportation; (vii) manufacturing; (viii) health (patients), and so forth.

The big picture of research is the interaction of random matrices, geometric functional analysis, and algorithms (theoretical computer science). We make the following observations:

- Random matrices are natural building blocks to model big data.
- At the heart of random matrix theory lies the realization that the *spectrum* of a random matrix **X** tends to stabilize as the dimensions of **X** grows to infinity.
- In the last few years, considerable progress was made on the more difficult local and nonasymptotic regimes. In the nonasymptotic regimes, the dimensions of **X** are **fixed** rather than grow to infinity.
- Connections among random matrix theory, quantum information theory, free probability, and statistics complete the picture.

The central objective of this section is to establish the fact that the circular law is the consequence of the more basic concept of "free entropy". Here we only sketch the key conceptual steps that complete the proof [1].

Circular and ring laws for eigenvalues are fundamental to random matrices. Non-Hermitian random matrices, and thus their eigenvalues, are complex values. See Chapter 6 for details. The circular law is observed for the (square) complex i.i.d. ensemble, while the ring law is for the rectangular complex i.i.d. ensemble. For an $N \times T$ complex matrix, the inner radius is $\sqrt{1-c}$, where $c = N/T \leq 1$. The circular law is the special case of the rectangular law for $N = T$ or $c = 1$.

The circular law [56] states that the empirical measure of the eigenvalues of a random $n \times n$ matrix, with i.i.d. entries of variance $1/n$, tends to the uniform law on the unit disc of the complex plane, as the dimension n tends to infinity. This universal result was proved rigorously by Tao and Vu [55], after 50 years of contributions. The circular law is universal, in the sense that it remains valid if one drops the Gaussian assumption of the entries of the matrix, while keeping the i.i.d. structure and the $1/n$ variance. The proof of this high dimensional phenomenon involves tools from potential theory, from additive combinatorics, and from asymptotic geometric analysis. The circular law phenomenon can be checked in the Gaussian case using the fact that the model is then exactly solvable. Actually, Ginibre has shown in the 1960s that if the entries are i.i.d.-centered complex Gaussians then the eigenvalues of such matrices form a Coulomb gas at temperature $1/n$ in dimension 2. This in turn suggests exploration of the analog of the circular law phenomenon in dimension ≥ 3, beyond random matrices. This led researchers to introduce in [57] stochastic interacting particle systems in which each particle is confined by an external field, and each pair of particles is subject to a singular repulsion. Under general assumptions and suitable scaling, the empirical measure of the particles converges, as the number of particles tends to infinity, to a probability measure that minimizes a natural energy-entropy functional. In the case of quadratic confinement and Coulomb repulsion, the limiting law is uniform on a ball.

Non-Hermitian matrices have a complex-valued eigenvalue distribution in general. In the Hermitian case, we work on the complex-valued matrix functions to search for real-valued eigenvalues, while we now have to work on a q-valued function to search complex-valued eigenvalues. See Table 1.1. For large non-Hermitian random matrices, we need quatartenionic free probability theory.

Boltzmann entropy (statistical physics), Shannon entropy (for classical information theory) and von Neumann entropy [39] (for quantum information) are all defined on the set of positive real-valued numbers. The eigenvalues of non-Hermitian random matrices are complex valued, in general. This basic fact suggests that the concepts and hence formulations based on Boltzmann entropy, Shannon entropy, and von Neumann entropy may not be sufficient for the theory of big data based on non-Hermitian random matrices.

Von Neumann entropy is defined as [58]

$$S(\rho) = \mathrm{Tr}\, \phi(\rho) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i$$

1 The similar justification of the ring law from a more basic concept is open at this point of writing.

Table 1.1 Comparison between classical, free, and quatartenionic free probability theories.

	Probability space	Algebra
Classical probability	Commutative	Commutative
Free probability	Noncommutative	Commutative
Quatartenionic free probability	Noncommutative	Noncommutative

Table 1.2 Comparison of different entropy definitions.

	Definition set	Mathematical expression	Remarks
Shannon/Boltzmann entropy	Positive real values	$S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i.$	p_i are positive
Von Neumann entropy	Positive real values	$S(\rho) = \mathrm{Tr}\, \phi(\rho) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i.$	λ_i are positive
Free entropy	Complex values	$\chi(\mu) := \iint \log\lvert x-y\rvert \mu(dx)\mu(dy).$	μ is complex on $\mathbb{C}$

where λ_i are the eigenvalues of ρ, a statistical operator, and $\phi : \mathbb{R}^+ \to \mathbb{R}$ is the continuous function $\phi(t) = -t \log t$. When studying non-Hermitian random matrices, we find that the eigenvalues λ_i are complex values, instead of real (positive) values. This suggests that von Neumann entropy is insufficient for the non-Hermitian data matrices. It is well known that Shannon entropy may be viewed as a special case of von Neumann entropy.

1.4.1 Boltzmann Entropy and H-Theorem

Consider a system of n distinguishable particles, each of them being in one of r possible states (typically energy levels). We have $n = n_1 + \cdots + n_r$ where n_i is the number of particles in state i. $S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$, where $\mathbf{p} := (p_1, \ldots, p_r)$. The quantity $S(p)$ is the Boltzmann entropy of the discrete probability distribution p. It appears here as an asymptotic additive degree of freedom per particle in a system with an infinite number of particles, each of them being in one of the r possible states, with population frequencies $p_1, ..., p_r$.

Returning to the motivations of Boltzmann, let us recall that the first principle of Carnot–Clausius thermodynamics states that the internal energy of an isolated system is constant, and the second principle states that there exists an extensive state variable called the entropy that can never decrease for an isolated system. Boltzmann wanted to derive the second principle from the idea (controversial, at that time) that matter is made with atoms. The H-theorem states that the entropy $S = -H$ is *monotonic* along the Boltzmann equation.

1.4.2 Shannon Entropy and Classical Information Theory

Boltzmann entropy also plays a fundamental role in communication theory [59]. It was founded in the 1940s by Claude Elwood Shannon (1916–2001) at Bell Labs, where it is known as "Shannon entropy."

> My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." (Claude E. Shannon, 1961)

1.4.3 Dan-Virgil Voiculescu and Free Central Limit Theorem

Free probability theory was forged in the 1980s by Dan-Virgil Voiculescu (1946–), while working on isomorphism problems in von Neumann operator algebras of free groups. Voiculescu discovered in the 1990s that free probability is the algebraic structure that appears naturally in the asymptotic global spectral analysis of random-matrix models as the dimension tends to infinity. Free probability theory comes with algebraic analogs of the central limit theorem and the Boltzmann entropy.

For the $n \times n$ complex matrix $\mathbf{A} \in \mathcal{M}_n(\mathbb{C})$, τ appears as an expectation with respect to the empirical spectral distribution. Denoting $\lambda_1(\mathbf{A}), \ldots, \lambda_n(\mathbf{A}) \in \mathbb{C}$ the eigenvalues of $\mathbf{A}$, we have

$$\tau(\mathbf{A}) = \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_k(\mathbf{A})} = \int x \mu_{\mathbf{A}}(dx), \text{ where } \mu_{\mathbf{A}} := \frac{1}{n}\sum_{k=1}^{n} \delta_{\lambda_k(\mathbf{A})}$$

We also obtain

$$\begin{aligned} 2\tau\left(\log\left((\mathbf{A}-z\mathbf{I})(\mathbf{A}-z\mathbf{I})^*\right)\right) &= \frac{1}{n}\log\left|\det\left(\mathbf{A}-z\mathbf{I}_n\right)\right| \\ &= \int \log|z-\lambda|\, d\mu_{\mathbf{A}}(\lambda) \\ &= \left(\log|z-\cdot| * \mu_{\mathbf{A}}\right)(z) \\ &=: -U_{\mu_{\mathbf{A}}}(z) \end{aligned}$$

The quantity $U_{\mu_{\mathbf{A}}}(z)$ is exactly the logarithmic potential at point $z \in \mathbb{C}$ of the probability measure $\mu_{\mathbf{A}}$.

Since $-\frac{1}{2\pi}\log|z-\cdot|$ is the so-called fundamental solution of the Laplace equation in dimension 2, it follows that, in the sense of Schwartz–Sobolev distributions, $\mu_{\mathbf{A}} = \frac{1}{2\pi}\Delta U_{\mu_{\mathbf{A}}}$. It is amazing to point out that the (discrete) empirical spectral distribution follows a (continuous) partial differential equation—the Laplace equation.

1.4.4 Free Entropy

Inspired by Boltzmann and Shannon on the central limit theory (CLT) of classical probability theory, we may ask if there exists, in free probability theory, a free entropy

functional, maximized by the semicircle law at fixed second moment, and which is monotonic along the free CLT.

The semicircle law is, for the free entropy, the analog of the Gaussian law for the Boltzmann entropy. The semicircle law on $[-2, 2]$ is the unique law that maximizes the Voiculescu entropy χ among the laws on $\mathbb{R}$ with a second moment equal to 1, for $\text{supp}(\mu) \subset \mathbb{R}$,

$$\arg\max\left\{\chi(\mu) : \int x^2\mu(dx) = 1\right\} = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{[-2,2]}(x)\,dx$$

How about laws on $\mathbb{C}$ (complex values) instead of $\mathbb{R}$ (real values)? When μ is a probability measure on $\mathbb{C}$, we will denote the Voiculescu entropy functional as

$$\chi(\mu) := \iint \log|x-y|\,\mu(dx)\mu\left(dy\right)$$

The uniform law on the unit disc is the unique law that maximizes the functional χ among the set of laws on $\mathbb{C}$ with the second moment (mean squared modulus) equal to 1 (here $z = x + iy$, and $dz = dxdy$) for $\text{supp}(\mu) \subset \mathbb{C}$

$$\arg\max\left\{\chi(\mu) : \int |z|^2\mu(dz) = 1\right\} = \frac{1}{\pi}\mathbf{1}_{\{z\in\mathbb{C}:|z|=1\}}dz$$

This phenomenon is known as the circular law. Under the uniform law on the unit disc, the real and the imaginary parts follow the semicircle law on $[-1, 1]$, and are not independent.

If one starts with a Hermitian random Gaussian matrix, the Gaussian unitary ensemble (GUE), then the same analysis is available, and produces a convergence to the semicircle law on $[-2, 2]$.

It turns out that the Voiculescu free entropy χ is monotonic along the Voiculescu free CLT. The Boltzmann–Shannon H-theorem interpretation of the CLT is thus remarkably valid in classical probability theory, and in free probability theory.

$\mathbf{A}$ and $\mathbf{B}$ are two $n \times n$ Hermitian matrices such that $\mu_{\mathbf{A}} \to \mu_a$, and $\mu_{\mathbf{B}} \to \mu_b$, in the sense of moments as $n \to \infty$, where μ_a and μ_b are two compactly supported laws on $\mathbb{R}$. Let $\mathbf{U}$ and $\mathbf{V}$ be independent random unitary matrices uniformly distributed on the unitary group (we say Haar unitary). Then

$$\mathbb{E}\mu_{\mathbf{UAU}^*+\mathbf{VBV}^*} \xrightarrow[n\to\infty]{*} \mu_a \boxplus \mu_b$$

This asymptotic freeness reveals that free probability is the algebraic structure that emerges from asymptotic analysis of large dimensional unitary invariant models of random matrices. As the functional χ is maximized by the uniform law on the unit disc, one may ask about an analog of the Wigner theorem for non-Hermitian random matrices. The answer is positive.

1.4.5 Jean Ginibre and his Ensemble of Non-Hermitian Random Matrices

The circular law for the Complex Ginibre ensemble, can be proved using the Voiculescu functional χ (maximized at fixed second moment by uniform law on unit disc). A simple model of random matrix is the Ginibre model:

$$\mathbf{G} = \begin{pmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \vdots & \vdots \\ G_{n1} & \cdots & G_{nn} \end{pmatrix}$$

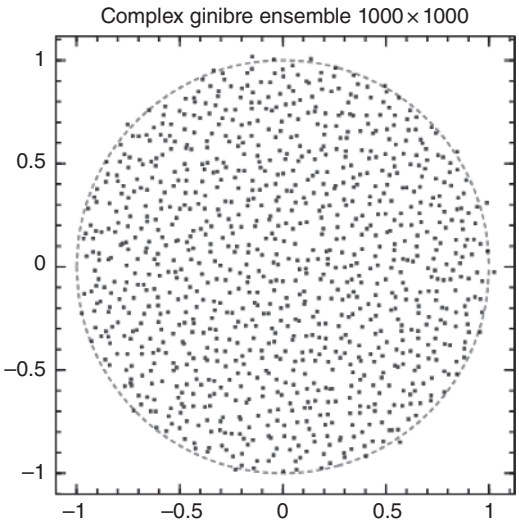

Figure 1.4 The eigenvalues of a single matrix drawn from the complex Ginibre ensemble of random matrices. The dashed line is the unit circle. This numerical experiment was performed using the promising Julia http://julialang.org/ (accessed August 17, 2016).

where $\left(G_{jk}\right)_{1\leqslant j,k\leqslant n}$ are i.i.d. random variables on $\mathbb{C}$, with $\operatorname{Re} G_{jk}, \operatorname{Im} G_{jk}$ of the Gaussian law of mean 0 and variance $1/(2n)$. The eigenvalues of a single matrix drawn from the complex Ginibre ensemble of random matrices is illustrated in Figure 1.4.

The density of $\mathbf{G}$ is proportional to

$$\prod_{j,k=1}^{n} \exp\left(-n\left|G_{jk}\right|^2\right) = \exp\left(-\sum_{j,k=1}^{n} n\left|G_{jk}\right|^2\right) = \exp\left(-n\operatorname{Tr}\left(\mathbf{G}\mathbf{G}^H\right)\right)$$

1.4.6 Circular Law for the Complex Ginibre Ensemble

The law of the eigenvalues is then proportional to

$$\exp\left(-n\sum_{j=1}^{n}\left|\lambda_j\right|^2\right)\prod_{1\leqslant j,k\leqslant n}\left|\lambda_j-\lambda_k\right|^2$$

This defines a determinantal process on $\mathbb{C}$: the complex Ginibre ensemble. In order to interpret the law of the eigenvalues as a Boltzmann measure, we put the Vandermonde determinant inside the exponential:

$$\exp\left(-n\sum_{j=1}^{n}\left|\lambda_j\right|^2+2\sum_{j<k}\log\left|\lambda_j-\lambda_k\right|\right)$$

If we encode the eigenvalues by the empirical measure

$$\mu_n := \frac{1}{n}\sum_{j=1}^{n}\delta_{\lambda_j}$$

this takes the form

$$e^{-n^2 I(\mu_n)}$$

where the "energy" $\mathcal{I}\left(\mu_n\right)$ of the configuration μ_n is defined via

$$\mathcal{I}\left(\mu_n\right) := \int |z|^2 d\mu\left(z\right) + \iint_{\neq} \log \frac{1}{|z - z'|} d\mu\left(z\right) d\mu\left(z'\right)$$

This suggests interpreting the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{G}$ as Coulomb gas of two-dimensional charged particles, confined by an an external field (quadratic potential) and subject to pair Coulomb repulsion.

$-\mathcal{I}$ can also be seen as a penalized Voiculescu functional. Minimizing a penalized functional is equivalent to minimizing without penalty but under constraint (Lagrange). Presently, if $\mathcal{M}$ is the set of probability measures on $\mathbb{C}$ then $\inf_{\mathcal{M}} \mathcal{I} > -\infty$ and the infimum is achieved at a *unique* probability measure μ_*, which is the **uniform** law on the unit disc of $\mathbb{C}$. The circular law is *universal*, in the sense that it remains valid if one drops the Gaussian assumption of the entries of the matrix, while keeping the i.i.d. structure and the $1/n$ variance.

How does the random discrete probability measure μ_n behave as $n \to \infty$? We may adopt a large deviations approach. Let $\mathcal{M}$ be the set of probability measures on $\mathbb{C}$. We may show that the functional $\mathcal{I} : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous for the topology of narrow convergence, is strictly convex, and has compact level sets. Let us consider a distance compatible with the topology. It can be shown that for every ball B for this distance

$$\mathbb{P}\left(\mu_n \in B\right) \approx \exp\left(-n^2 \inf_B \left(\mathcal{I} - \inf_B \mathcal{I}\right)\right)$$

The first Borel–Cantelli lemma allows one to deduce that almost surely

$$\lim_{n \to 0} \mu_n = \mu_* = \arg\inf \mathcal{I} = \frac{1}{\pi} \mathbf{1}_{\{z \in \mathbb{C} : |z| \leqslant 1\}} dz$$

where $z = x + iy$ and $dz = dxdy$. This phenomenon is known as the circular law. If one starts with a Hermitian random Gaussian matrix—the Gaussian unitary ensemble (GUE)—then the same analysis is available, and produces a convergence to the semicircle law on $[-2, 2]$.

1.5 Smart Grid

Roughly speaking, a smart grid can be viewed as two flows: (i) information, and (ii) electric power. The information flow is used for grid control. Communications, sensing, and control must be considered jointly. At an abstract level, the smart grid can be viewed as an "energy Internet." This is very relevant to the Internet of Things, for machine-to-machine communications.

The vision of a smart transmission grid is illustrated in Figure 1.5. As a roadmap for research and development, the smart features of the transmission grid are envisaged and summarized as digitization, flexibility, intelligence, resilience, sustainability, and customization. The enabling technologies include [60]:

- *New materials and alternative clean energy resources.* The high penetration of alternative clean energy resources will mitigate the conflicts between the development of human society and environment sustainability.

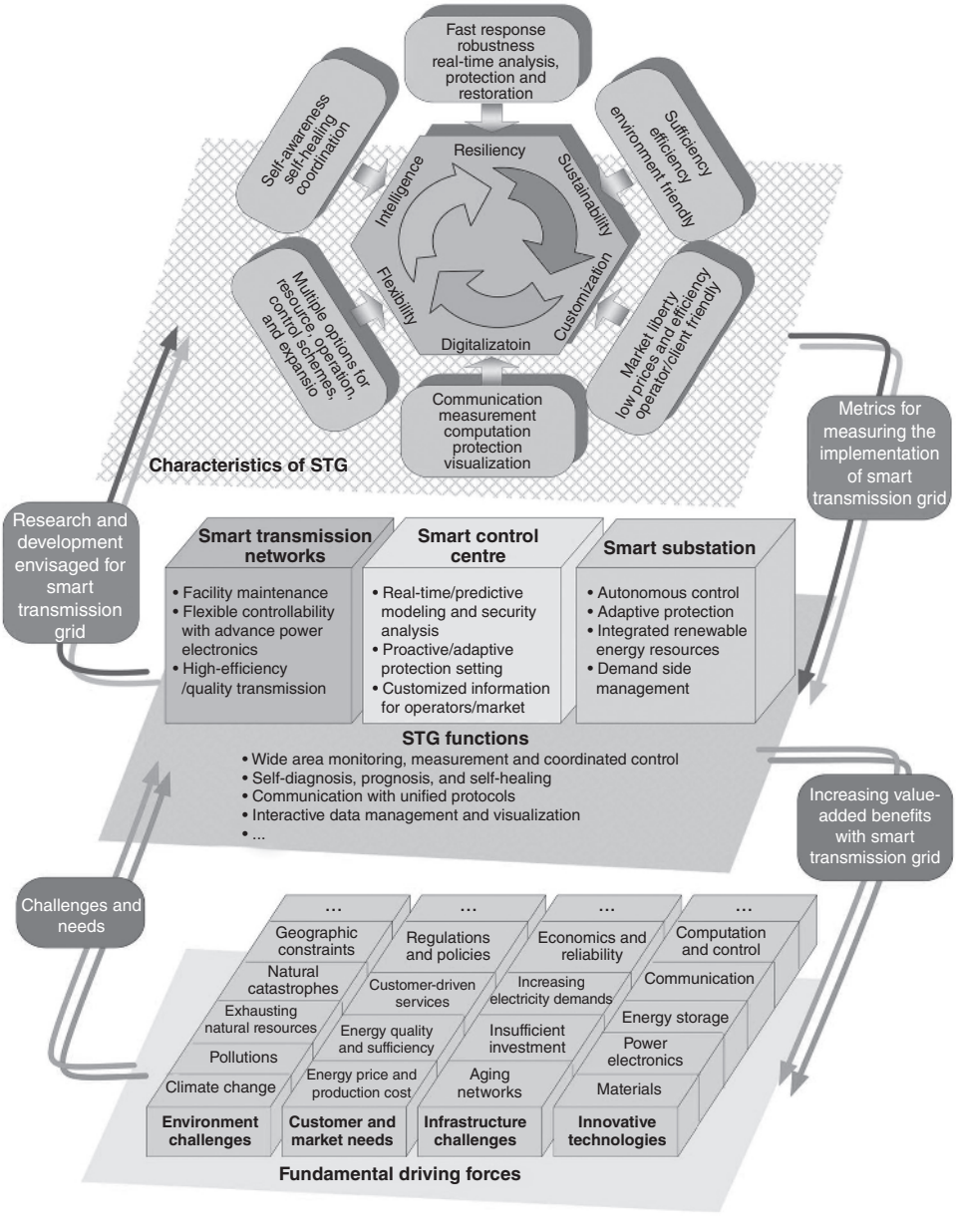

Figure 1.5 Vision of a smart transmission grid. Source: Reproduced from [60] with Permission of IEEE.

- *Advanced power electronics and devices.* These greatly improve the quality of power supply and flexibility of power flow control.
- *Sensing and measurement.* The basis for communications, computing, control, and intelligence.
- *Communications.* Adaptive communication networks will allow open-standardized communication protocols to operate on a unique platform. Real-time control based on fast and accurate information exchange on different platforms will improve system resilience by the enhancement of system reliability and security, and optimization of the transmission asset utilization.
- *Advanced computing and control methodologies*. High-performance computing, parallel, and distributed computing technologies will enable real-time modeling and simulation of complex power systems. The accuracy of the situation awareness will be improved for further suitable operations and control strategies. Advanced control methodologies and novel distributed control paradigms will be needed to automate the entire customer-centric power-delivery network.
- *Mature power market regulation and policies*. These improve the transparency, liberty, and competition of the power market. High customer interaction with the electricity consumption should be enabled and encouraged.
- *Intelligent technologies.* These enable fuzzy logic reasoning, and knowledge discovery.

1.6 Big Data and Smart Grid

Our knowledge is dominated by the scales in which our observations are made. Our slogan is "data is science and science is data." This book treats big data as the foundation for the smart grid, an approach that is consistent with [39] and [40]. In other words, the science of smart grid is a combination of distributed sensing and a distributed network with the electric power grid. See Chapter 11 for details about why big data should be tied together with the smart grid. The central task is to understand the statistical knowledge of the massive datasets and make sense of these data.

Large random matrices are used to model large datasets. It is our firm belief that large random matrices are the basic building blocks for our science. It is the calculus for data. From the point of view of probability and statistics, after living in the age of vector-valued random variables, we are entering a new age of big data, an age of matrix-valued random variables. Initially, Newton and Leibniz developed the calculus of $f(x)$, where x is a free variable. Later, we study $f(X)$ where is a scalar-valued random variable (a function defined on the sample space). Then, we study $f(\mathbf{x})$ where $\mathbf{x} = \left[X_1, \ldots, X_N\right]^T$ is a vector-valued random variable. Now we are entering an age of studying $f(\mathbf{X})$ where $\mathbf{X} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) \in \mathbb{C}^{N\times n}$ is a matrix-valued random variable. In particular, we are interested in the asymptotic regime of

$$N \to \infty, n \to \infty \text{ but } \frac{N}{n} \to c \in (0, \infty)$$

Alternatively, we are interested in the nonasymptotic regime where

N, n are large but finite

For a complex quantum system (a system with many degrees of freedom)—such as atoms, nuclei, fundamental particles, it is almost impossible to imagine a theory that

is exploitable enough to compute accurately, for instance, the energy levels of such a system. Antenna sensors, smart meters, PMUs, and stocks are analogies. The model of random particles [56] is relevant in this context, for example. Energy and entropy are two drivers.

1.7 Reading Guide

The core material of this book provides a comprehensive study of large random matrices for big data applications (Part I) (see Figure 1.6). After this has been accomplished, we make connections with selected smart grid applications (Part II) and selected applications in communications and sensing (Part III). Random matrix theory has been used as the unifying tool to tie the three parts together. Very often, connections are made at the mathematical level.

Missing links are, however, inevitably frequent because the majority of materials (90% we guess) appear, for the first time, in book form. Even worse, most materials are treated, for the first time, in the context of engineering applications. The main obstacle when reading this book is the mathematical depth. Although tested in the class room, the limited size of this book makes it impossible to present all the material in a self-contained manner.

For the large random matrix, the trick is to convert two-dimensional matrix problems into one-dimensional problems by using the eigenvalue distribution—forget about the eigenvectors for the moment. As a result, we can study the function of $f\left(\lambda_i\right), i = 1, \ldots, n$, for some function f. Various functions f are defined for different applications.

To interpret our empirical discovery in [61] we connected our results with quantum information theory; see [39] for this link. About 200 pages were also dedicated to random matrix theory in [39]. It was realized that our discovery was caused by the high dimensionality of the problem, which lead to the "concentration of measure" phenomenon, a high-dimensional effect, or a property of a large number of variables, for which functions with small local oscillations are almost constant. In this connection,

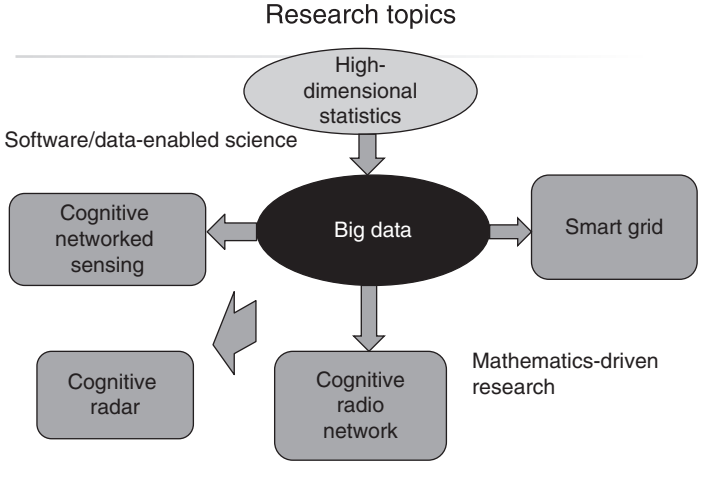

Figure 1.6 Big data vision.

the first author's book [40] was born. The current book can be viewed applying [39] and [40]. The use of random matrix theory as the unifying theme to model large wireless networks, smart grid and big data was explicitly pointed out in [39]. [40] was written to support this big vision. These three books are complementary. We closed the circle during the writing of three books. Now we are revisiting random matrix theory, with the emphasis on the latest results, which are also applicable to the problems we have in mind—smart grid and big data. During this adventure, the most remarkable experience with random matrix theory is our feeling of being shocked by its usefulness, beauty, depth and fertility, as pointed out in the preface of [62]. According to him, usefulness is usually measured by the utility of the topic outside mathematics. Beauty is a quality of much the material, but is often something only a trained eye can see. "Depth comes via the linking together of multiple ideas and topics, often seemingly removed from the original context. And fertility means that with a reasonable effort there are new results, some useful, some with beauty, and a few maybe with depth, still waiting to be found."

In Chapter 1, we started our book with some challenges for big data. Chapter 2 gives an overview of the framework for the mathematical framework needed for the analysis of big data. We use a bottom-up approach to lay the foundations using large random matrices to summarize the large datasets (big data).

Chapter 3 gives the fundamentals of large dimensional random matrices. One motivation is to model the large datasets using large random matrices. It is our belief that large dimensional random matrices are the foundation for the analysis of big data; this chapter is the basic material for next-generation engineers and researchers.

Chapter 4, by studying the central limit theory for linear spectral statistics, addresses the spectral analysis of large dimensional random matrices. The main reason is because many important statistics in multivariate statistical analysis can be expressed as functionals of the empirical spectral distribution of some random matrices.

Chapter 5 studies the Hermtian free probability theory. The idea of exploiting "large models" is the unified theme of this whole book. As a result, large random matrices are natural building blocks for the entire theoretical framework. As large random matrices can be regarded as free random variables, matrix-valued free probability is discussed to study the variables.

Chapter 6 studies (large) non-Hermitian random matrices using the newly developed quatartenionic free probability theory. Most results appear in book form for the first time.

Chapter 7 deals with data collection. Data storage is central to big data. For many applications, we often cannot afford the luxury of saving all the raw data generated by the system (or network) for future processing. One fundamental challenge is to choose which types of information are stored. As we deal with streaming data, real-time processing is required.

Chapter 8 deals with anomaly detection using large random matrices. One objective is is to study the denial of service using big data. We understand how the large data size affects the matrix hypothesis detection.

Requirements for applying big data to smart grid are addressed in Chapter 9. The technical challenges are discussed in Chapter 10. And big data topics for smart grid are addressed in Chapter 11.

Chapter 12 introduces grid monitoring and state estimation using phasor measurement units (PMUs). Chapter 13 gives an exhaustive treatment of false data injection

attacks in the context of state estimation. It is well known that cyber security is the most important task facing engineers and researchers. We use false data injection to attack against state estimation.

Chapter 14 briefly discusses the demand response.

Chapter 15 addresses communications topics for smart grids. To control the power grid we need sensing and communications to tie together the whole grid. High-performance computing and distributed computing are two enablers.

Bibliographical Remarks

This current book together with another two books [39,40] pursues a paradigm of modeling big data using large random matrices. To the best of our knowledge, this vision was explicitly spelled out and formulated analytically for the first time in November of 2011 during the writing of [39].

Section 1.1.5 draws on material from [4, 26, 27]. We are now facing the data deluge. [63]. For big data we follow [22], which is an excellent review and tutorial. Labrinidis and Jagadish (2012) [24] is very insightful. We have followed [24] for insights.

The state-of-the-art of big data is that there is no clear definition for big data, or the adopted theoretical framework. Our aim of these three books is to attempt to define our big data problems in a random matrix way. There is no claim of solving all big data problems using one framework. In Section 1.3 (Definition 1.3.1), we use three conditions to define our problems related to big data. We limit the potential applications of our methods using Definition 1.3.1. Clarity and rigor, on the other hand, are achieved. Our three books are aimed at addressing the consequences of Definition 1.3.1 in the context of large random matrices.

We have drawn from [45] for some parts of Section 1.2. For Example 1.2.3, we also drew from [47]. Challenges include: (i) Real-time processing [64] is challenging; (ii) other technical challenges [65]; (iii) signal processing [66].

Example 1.3.2 and Example 1.3.3 are adapted from [49, 67].

We adopt statistical methods to study big data. Fisher [68] states that the purpose of statistical methods is to reduce a large quantity of data to a small amount of data that is capable of containing as much of the relevant information as possible in the original data. Because the data will generally supply a large number of "facts," many more than are sought, much information in the data is irrelevant. This brings to the fore the Fisherial dictum [69, p. 1] that statistical analysis via the reduction of the data is the process of extracting the irrelevant information. This may be accomplished by modeling a hypothetical population specified by relative few parameters. See [44,70] for one application in modeling big data in large wireless networks.

Functions are the core for the practical applications. Tao's excellent text [67] relies heavily on Talagrand's concentration theorem for convex functions. High-dimensional spaces were used to model big data, originally in [39] and then in [40], by using large random matrices.

In Example 1.3.7, we draw some material from [71]. Example 1.3.8 takes results from [72–75]. More recent work is [76–78]. In [44,70] we used non-Hermitian random matrices to model big data collected in a large-scale wireless cognitive radio network. We follow [54] in Example 1.3.9.

Example 1.3.10 follows [79].

Example 1.3.11 follows [80].

Example 1.3.12 is taken from [81].

Example 1.3.16 follows [56, 57].

In Section 1.4 we follow [56] for the development of the unified mathematical theory for big data.