Preface

Learning directly from data is critical to a host of disciplines in engineering and the physical, social, and life sciences. Modern society is literally driven by an interconnected web of data exchanges at rates unseen before, and it relies heavily on decisions inferred from patterns in data. There is nothing fundamentally wrong with this approach, except that the inference and learning methodologies need to be anchored on solid foundations, be fair and reliable in their conclusions, and be robust to unwarranted imperfections and malicious interference.

P.1 EMPHASIS ON FOUNDATIONS

Given the explosive interest in data-driven learning methods, it is not uncommon to encounter claims of superior designs in the literature that are substantiated mainly by sporadic simulations and the potential for "life-changing" applications rather than by an approach that is founded on the well-tested scientific principle to inquiry. For this reason, one of the main objectives of this text is to highlight, in a unified and formal manner, the firm mathematical and statistical pillars that underlie many popular data-driven learning and inference methods. This is a nontrivial task given the wide scope of techniques that exist, and which have often been motivated independently of each other. It is nevertheless important for practitioners and researchers alike to remain cognizant of the common foundational threads that run across these methods. It is also imperative that progress in the domain remains grounded on firm theory. As the aphorism often attributed to Lewin (1945) states, "*there is nothing more practical than a good theory*." According to Bedeian (2016), this saying has an even older history.

Rigorous data analysis, and conclusions derived from experimentation and theory, have been driving science since time immemorial. As reported by Heath (1912), the Greek scientist Archimedes of Syracuse devised the now famous Archimedes' Principle about the volume displaced by an immersed object from observing how the level of water in a tub rose when he sat in it. In the account by Hall (1970), Gauss' formulation of the least-squares problem was driven by his desire to predict the future location of the planetoid Ceres from observations of its location over 41 prior days. There are numerous similar examples by notable scientists where experimentation led to hypotheses and from there to substantiated theories and well-founded design methodologies. Science is also full of progress in the reverse direction, where theories have been developed first

to be validated only decades later through experimentation and data analysis. Einstein (1916) postulated the existence of gravitational waves over 100 years ago. It took until 2016 to detect them! Regardless of which direction one follows, experimentation to theory or the reverse, the match between solid theory and rigorous data analysis has enabled science and humanity to march confidently towards the immense progress that permeates our modern world today.

For similar reasons, data-driven learning and inference should be developed with strong theoretical guarantees. Otherwise, the confidence in their reliability can be shaken if there is over-reliance on "proof by simulation or experience." Whenever possible, we explain the underlying models and statistical theories for a large number of methods covered in this text. A good grasp of these theories will enable practitioners and researchers to devise variations with greater mastery. We weave through the foundations in a coherent and cohesive manner, and show how the various methods blend together techniques that may appear decoupled but are actually facets of the same common methodology. In this process, we discover that a good number of techniques are well-grounded and meet proven performance guarantees, while other methods are driven by ingenious insights but lack solid justifications and cannot be guaranteed to be "fail-proof."

Researchers on learning and inference methods are of course aware of the limitations of some of their approaches, so much so that we encounter today many studies, for example, on the topic of "explainable machine learning." The objective here is to understand why learning algorithms produce certain recommendations. While this is an important area of inquiry, it nevertheless highlights one interesting shift in paradigm. In the past, the emphasis would have been on designing inference methods that respond to the input data in certain desirable and controllable ways. Today, in many instances, the emphasis is to stick to the available algorithms (often, out of convenience) and try to understand or explain why they are responding in certain ways to the input!

Writing this text has been a rewarding journey that took me from the early days of statistical mathematical theory to the modern state of affairs in learning theory. One can only stand in awe at the wondrous ideas that have been introduced by notable researchers along this trajectory. At the same time, one observes with some concern an emerging trend in recent years where solid foundations receive less attention in lieu of "speed publishing" and over-reliance on "illustration by simulation." This is of course not the norm and most researchers in the field stay honest to the scientific approach to inquiry and design. After concluding this comprehensive text, I stand humbled at the realization of "*how little we know*!" There are countless questions that remain open, and even for many of the questions that have been answered, their answers rely on assumptions or (over)simplifications. It is understandable that the complexity of the problems we face today has increased manifold, and ingenious approximations become necessary to enable tractable solutions.

P.2 GLIMPSE OF HISTORY

Reading through the text, the alert reader will quickly realize that the core foundations of modern-day machine learning, data analytics, and inference methods date back for at least two centuries, with contributions arising from a range of fields including mathematics, statistics, optimization theory, information theory, signal processing, communications, control, and computer science. For the benefit of the reader, I reproduce here with permission from IEEE some historical remarks from the editorial I published in Sayed (2018). I explained there that these disciplines have generated a string of "big ideas" that are driving today multi-faceted efforts in the age of "big data" and machine learning. Generations of students in the statistical sciences and engineering have been trained in the art of modeling, problem solving, and optimization. Their algorithms power everything from cell phones, to spacecraft, robotic explorers, imaging devices, automated systems, computing machines, and also recommender systems. These students mastered the foundations of their fields and have been well prepared to contribute to the growth of data analysis and machine learning solutions.

As the list below shows, many well-known engineering and statistical methods have actually been motivated by data-driven inquiries, even from times remote. The list is a tour of some older historical contributions, which is of course biased by my personal preferences and is not intended to be exhaustive. It is only meant to illustrate how concepts from statistics and the information sciences have always been at the center of promoting big ideas for data and machine learning. Readers will encounter these concepts in various chapters in the text. Readers will also encounter additional historical accounts in the concluding remarks of each chapter, and in particular comments on newer contributions and contributors.

Let me start with Gauss himself, who in 1795 at the young age of 18, was fitting lines and hyperplanes to astronomical data and invented the least-squares criterion for regression analysis — see the collection of his works in Gauss (1903). He even devised the recursive least-squares solution to address what was a "big" data problem for him at the time: He had to avoid tedious repeated calculations by hand as more observational data became available. What a wonderful big idea for a data-driven problem! Of course, Gauss had many other big ideas.

de Moivre (1733), Laplace (1812), and Lyapunov (1901) worked on the central limit theorem. The theorem deals with the limiting distribution of averages of "large" amounts of data. The result is also related to the law of "large" numbers, which even has the qualification "large" in its name. Again, big ideas motivated by "large" data problems.

Bayes (ca mid 1750s) and Laplace (1774) appear to have independently discovered the Bayes rule, which updates probabilities conditioned on observations — see the article by Bayes and Price (1763). The rule forms the backbone of much of statistical signal analysis, Bayes classifiers, Naïve classifiers, and Bayesian networks. Again, a big idea for data-driven inference.

Fourier (1822), whose tools are at the core of disciplines in the information sciences, developed the phenomenal Fourier representation for signals. It is meant to transform data from one domain to another to facilitate the extraction and visualization of information. A big transformative idea for data.

Forward to modern times. The fast Fourier transform (FFT) is another example of an algorithm driven by challenges posed by data size. Its modern version is due to Cooley and Tukey (1965). Their algorithm revolutionized the field of discrete-time signal processing, and FFT processors have become common components in many modern electronic devices. Even Gauss had a role to play here, having proposed an early version of the algorithm some 160 years before, again motivated by a data-driven problem while trying to fit astronomical data onto trigonometric polynomials. A big idea for a data-driven problem.

Closer to the core of statistical mathematical theory, both Kolmogorov (1939) and Wiener (1942) laid out the foundations of modern statistical signal analysis and optimal prediction methods. Their theories taught us how to extract information optimally from data, leading to further refinements by Wiener's student Levinson (1947) and more dramatically by Kalman (1960). The innovations approach by Kailath (1968) exploited to great effect the concept of orthogonalization of the data and recursive constructions. The Kalman filter is applied across many domains today, including in financial analysis of market data. Kalman's work was an outgrowth of the model-based approach to system theory advanced by Zadeh (1954). The concept of a recursive solution from streaming data was a novelty in Kalman's filter; the same concept is commonplace today in online learning techniques. Again, big ideas for recursive inference from data.

Cauchy (1847) early on, and Robbins and Monro (1951) a century later, developed the powerful gradient descent method for root finding, which is also recursive in nature. Their techniques have grown to motivate huge advances in stochastic approximation theory. Notable contributions that followed include the work by Rosenblatt (1957) on the Perceptron algorithm for single-layer networks, and the impactful delta rule by Widrow and Hoff (1960), widely known as the LMS algorithm in the signal processing literature. Subsequent work on multi-layer neural networks grew out of the desire to increase the approximation power of single-layer networks, culminating with the backpropagation method of Werbos (1974). Many of these techniques form the backbone of modern learning algorithms. Again, big ideas for recursive online learning.

Shannon (1948a,b) contributed fundamental insights to data representation, sampling, coding, and communications. His concepts of entropy and information measure helped quantify the amount of uncertainty in data and are used, among other areas, in the design of decision trees for classification purposes and in deriving learning algorithms for neural networks. Nyquist (1928) contributed to the understanding of data representations as well. Big ideas for data sampling and data manipulation.

Bellman (1957a,b), a towering system-theorist, introduced dynamic programming and the notion of the curse of dimensionality, both of which are core un-

derpinnings of many results in learning theory, reinforcement learning, and the theory of Markov decision processes. Viterbi's algorithm (1967) is one notable example of a dynamic programming solution, which has revolutionized communications and has also found applications in hidden Markov models widely used in speech recognition nowadays. Big ideas for conquering complex data problems by dividing them into simpler problems.

Kernel methods, building on foundational results by Mercer (1909) and Aronszajn (1950), have found widespread applications in learning theory since the mid 1960s with the introduction of the kernel Perceptron algorithm. They have also been widely used in estimation theory by Parzen (1962), Kailath (1971), and others. Again, a big idea for learning from data.

Pearson and Fisher launched the modern field of mathematical statistical signal analysis with the introduction of methods such as principal component analysis (PCA) by Pearson (1901) and maximum likelihood and linear discriminant analysis by Fisher (1912,1922,1925). These methods are at the core of statistical signal processing. Pearson (1894,1896) also had one of the earliest studies of fitting a mixture of Gaussian models to biological data. Mixture models have now become an important tool in modern learning algorithms. Big ideas for data-driven inference.

Markov (1913) introduced the formalism of Markov chains, which is widely used today as a powerful modeling tool in a variety of fields including word and speech recognition, handwriting recognition, natural language processing, spam filtering, gene analysis, and web search. Markov chains are also used in Google's PageRank algorithm. Markov's motivation was to study letter patterns in texts. He laboriously went through the first 20,000 letters of a classical Russian novel and counted pairs of vowels, consonants, vowels followed by a consonant, and consonants followed by a vowel. A "big" data problem for his time. Great ideas (and great patience) for data-driven inquiries.

And the list goes on, with many modern day and ongoing contributions by statisticians, engineers, and computer scientists to network science, distributed processing, compressed sensing, randomized algorithms, optimization, multi-agent systems, intelligent systems, computational imaging, speech processing, forensics, computer visions, privacy and security, and so forth. We provide additional historical accounts about these contributions and contributors at the end of the chapters.

P.3 ORGANIZATION OF THE TEXT

The text is organized into three volumes, with a sizable number of problems and solved examples. The table of contents provides details on what is covered in each volume. Here we provide a condensed summary listing the three main themes:

1. (**Volume I: Foundations**). The first volume covers the *foundations* needed for a solid grasp of inference and learning methods. Many important topics are covered in this part, in a manner that prepares readers for the study of inference and learning methods in the second and third volumes. Topics include: matrix theory, linear algebra, random variables, Gaussian and exponential distributions, entropy and divergence, Lipschitz conditions, convexity, convex optimization, proximal operators, gradient-descent, mirror-descent, conjugate-gradient, subgradient methods, stochastic optimization, adaptive gradient methods, variance-reduced methods, distributed optimization, and nonconvex optimization. Interestingly enough, the following concepts occur time and again in all three volumes and the reader is well-advised to develop familiarity with them: convexity, sample mean and law of large numbers, Gaussianity, Bayes rule, entropy, Kullback-Leibler divergence, gradient-descent, least squares, regularization, and maximum-likelihood. The last three concepts are discussed in the initial chapters of the second volume.

2. (**Volume II: Inference**). The second volume covers inference methods. By "inference" we mean techniques that infer some unknown variable or quantity from observations. The difference we make between "inference" and "learning" in our treatment is that inference methods will target situations where some prior information is known about the underlying signal models or signal distributions (such as their joint probability density functions or generative models). The performance by many of these inference methods will be the ultimate goal that learning algorithms, studied in the third volume, will attempt to emulate. Topics covered here include: mean-square-error inference, Bayesian inference, maximum-likelihood estimation, expectation maximization, expectation propagation, Kalman filters, particle filters, posterior modeling and prediction, Markov Chain Monte Carlo methods, sampling methods, variational inference, latent Dirichlet allocation, hidden Markov models, independent component analysis, Bayesian networks, inference over directed and undirected graphs, Markov decision processes, dynamic programming, and reinforcement learning.

3. (**Volume III: Learning**). The third volume covers learning methods. Here, again, we are interested in inferring some unknown variable or quantity from observations. The difference, however, is that the inference will now be solely data-driven, i.e., based on available data and not on any assumed knowledge about signal distributions or models. The designer is only given a collection of observations that arise from the underlying (unknown) distribution. New phenomena arise related to generalization power, overfitting, and underfitting depending on how representative the data is and how complex or simple the approximate models are. The target is to use the data to learn about the quantity of interest (its value or evolution). Topics covered here include: least-squares methods, regularization, nearest-neighbor rule, self-organizing maps, decision trees, naïve Bayes classifier, linear discrimi-

nant analysis, principal component analysis, dictionary learning, Perceptron, support vector machines, bagging and boosting, kernel methods, Gaussian processes, generalization theory, feedforward neural networks, deep belief networks, convolutional networks, generative networks, recurrent networks, explainable learning, adversarial attacks, and meta learning.

Figure P.1 shows how various topics are grouped together in the text; the numbers in the boxes indicate the chapters where these subjects are covered. The figure can be read as follows. For example, instructors wishing to cover:

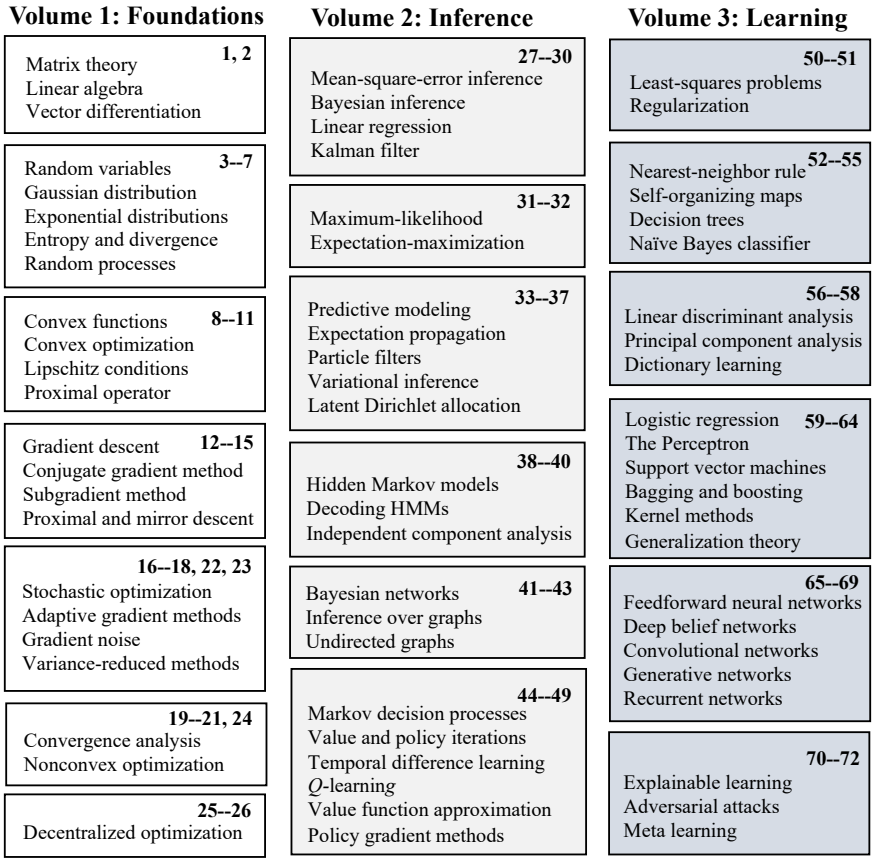

Figure P.1 Organization of the text.

(a) Background material on linear algebra and matrix theory: they can use Chapters 1 and 2.

(b) Background material on random variables and probability theory: they can select from Chapters 3 through 7.

(c) Background material on convex functions and convex optimization: they can use Chapters 8 through 11.

The three groupings **(a)–(c)** contain introductory core concepts that are needed for subsequent chapters. For instance, instructors wishing to cover gradient descent and iterative optimization techniques, would then proceed to Chapters 12 through 15, while instructors wishing to cover stochastic optimization methods would use Chapters 16-24 and so forth. Figure P.2 provides a representation of the estimated dependencies among the chapters in the text. The chapters are color-coded depending on the volume they appear in. An arrow from Chapter a towards Chapter b implies that the material in the latter chapter benefits from the material in the earlier chapter. In principle, we should have added arrows from Chapter 1, which covers background material on matrix and linear algebra, into all other chapters. We ignored obvious links of this type to avoid crowding the figure.

P.4 HOW TO USE THE TEXT

Each chapter in the text consists of several blocks: **(1)** the main text where theory and results are presented, **(2)** a couple of solved examples to illustrate the main ideas and also to extend them, **(3)** comments at the end of the chapter providing a historical perspective and linking the references through a motivated timeline, **(4)** a list of problems of varying complexity, **(5)** appendices when necessary to cover some derivations or additional topics, and **(6)** references. In total, there are close to 470 solved examples and 1350 problems in the text. *A solutions manual is available to instructors.*

In the comments at the end of each chapter I list in boldface the life span of some influential scientists whose contributions have impacted the results discussed in the chapter. The dates of birth and death rely on several sources, including the MacTutor History of Mathematics Archive, Encyclopedia Britannica, Wikipedia, Porter and Ogilvie (2000), and Daintith (2008).

Several of the solved examples in the text involve computer simulations on datasets to illustrate the conclusions. The simulations, and several of the corresponding figures, were generated using the software program Matlab©, which is a registered trademark of MathWorks Inc., 24 Prime Park Way, Natick, MA 01760-1500, `www.mathworks.com`. The computer codes used to generate the figures are provided "as is" and without any guarantees. While these codes are useful for the instructional purposes of the book, they are not intended to be examples of full-blown or optimized designs; practitioners should use them at their own risk. We have made no attempts to optimize the codes, perfect them, or even check them for absolute accuracy. On the contrary, in order to keep the codes at a level that is easy to follow by students, we have often decided to sacrifice performance or even programming elegance in lieu of simplicity. Students can use the computer codes to run variations of the examples shown in the text.

In principle, each volume could serve as the basis for a master-level graduate

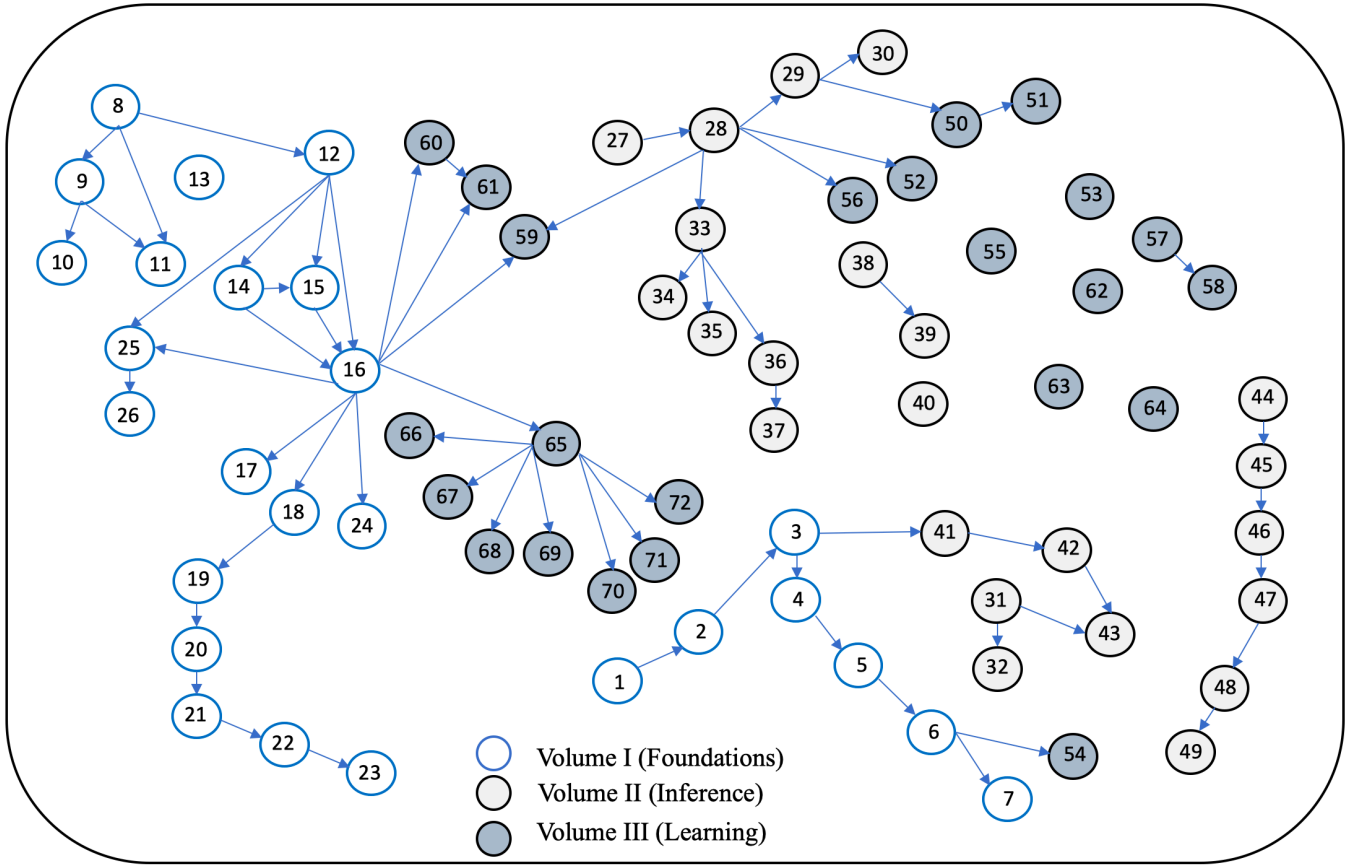

Figure P.2 A diagram showing the approximate dependencies among the chapters in the text. The color scheme identifies chapters from the same volume, with the numbers inside the circles referring to the chapter numbers.

course, such as courses on *Foundations of Data Science* (volume I), *Inference from Data* (volume II), and *Learning from Data* (volume III). Once students master the foundational concepts covered in volume I (especially in Chapters 1-17), they will be able to grasp the topics from volumes II and III more confidently. Instructors need not cover volumes II and III in this sequence; the order can be switched depending on whether they desire to emphasize data-based learning over model-based inference or the reverse. Depending on the duration of each course, one can also consider covering subsets of each volume by focusing on particular subjects. The following grouping explains how chapters from the three volumes cover specific topics and could be used as reference material for several potential course offerings:

1. (**Core foundations, Chapters 1–11, Vol. I**): matrix theory, linear algebra, random variables, Gaussian and exponential distributions, entropy and divergence, Lipschitz conditions, convexity, convex optimization, and proximal operators. These chapters can serve as the basis for an introductory course on foundational concepts for mastering data science.
2. (**Stochastic optimization, Chapters 12–26, Vol. I**): gradient-descent, mirror-descent, conjugate-gradient, subgradient methods, stochastic optimization, adaptive gradient methods, variance-reduced methods, convergence analyses, distributed optimization, and nonconvex optimization. These chapters can serve as the basis for a course on stochastic optimization for both convex and non-convex environments, with attention to performance and convergence analyses. Stochastic optimization is at the core of most modern learning techniques, and students will benefit greatly from a solid grasp of this topic.
3. (**Statistical or Bayesian inference, Chapters 27–37, 40, Vol. II**): mean-square-error inference, Bayesian inference, maximum-likelihood estimation, expectation maximization, expectation propagation, Kalman filters, particle filters, posterior modeling and prediction, Markov Chain Monte Carlo methods, sampling methods, variational inference, latent Dirichlet allocation, and independent component analysis. These chapters introduce students to optimal methods to extract information from data, under the assumption that the underlying probability distributions or models are known. In a sense, these chapters reveal limits of performance that future data-based learning methods, covered in subsequent chapters, will try to emulate when the models are not known.
4. (**Probabilistic graphical models, Chapters 38, 39, 41–43, Vol. II**): hidden Markov models, Bayesian networks, inference over directed and undirected graphs, factor graphs, message passing, belief propagation, and graph learning. These chapters can serve as the basis for a course on Bayesian inference over graphs. Several methods and techniques are discussed along with supporting examples and algorithms.

5. (**Reinforcement learning, Chapters 44–49, Vol. II**): Markov decision processes, dynamic programming, value and policy iterations, temporal difference learning, $Q-$learning, value function approximation, and policy gradient methods. These chapters can serve as the basis for a course on reinforcement learning. They cover many relevant techniques, illustrated by means of examples, and include performance and convergence analyses.

6. (**Data-driven and online learning, Chapters 50–64, Vol. III**): least-squares methods, regularization, nearest-neighbor rule, self-organizing maps, decision trees, naïve Bayes classifier, linear discriminant analysis, principal component analysis, dictionary learning, Perceptron, support vector machines, bagging and boosting, kernel methods, Gaussian processes, and generalization theory. These chapters cover a variety of methods for learning directly from data, including various methods for online learning from sequential data. The chapters also cover performance guarantees from statistical learning theory.

7. (**Neural networks, Chapters 65–72, Vol. III**): feedforward neural networks, deep belief networks, convolutional networks, generative networks, recurrent networks, explainable learning, adversarial attacks, and meta learning. These chapters cover various architectures for neural networks and the respective algorithms for training them. The chapters also cover topics related to explainability and adversarial behavior over these networks.

The above groupings assume that students have been introduced to background material on matrix theory, random variables, entropy, convexity, and gradient-descent methods. One can, however, rearrange the groupings by designing stand-alone courses where the background material is included along with the other relevant chapters. By doing so, it is possible to devise various course offerings, covering themes such as stochastic optimization, online or sequential learning, probabilistic graphical models, reinforcement learning, neural networks, Bayesian machine learning, kernel methods, decentralized optimization, and so forth. Figure P.3 shows several suggested selections of topics from across the text, and the respective chapters, which can be used to construct courses with particular emphasis. Other selections are of course possible, depending on individual preferences and on the intended breadth and depth for the courses.

P.5 SIMULATION DATASETS

In several examples in this work we run simulations that rely on publicly available datasets. The sources for these datasets are acknowledged in the appropriate locations in the text. Here we provide an aggregate summary for ease of reference:

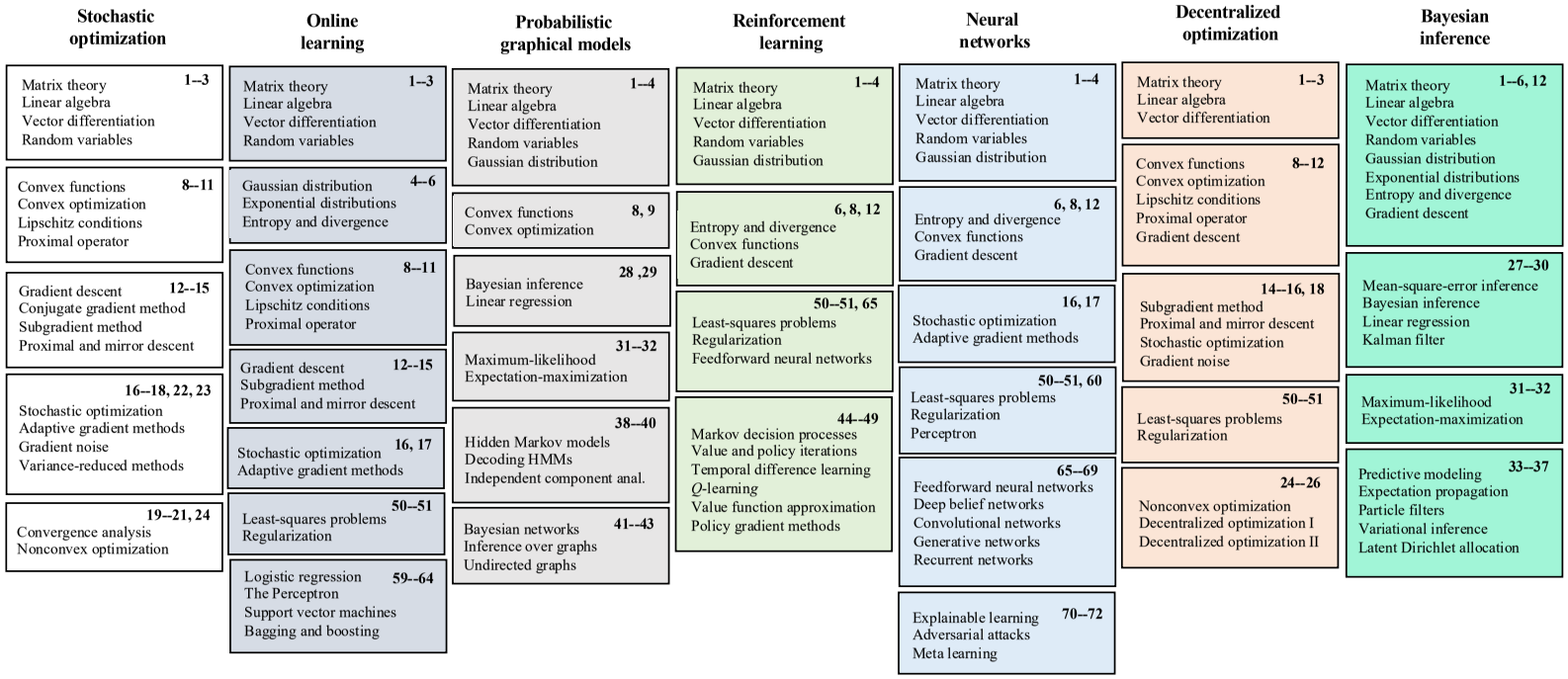

Figure P.3 Suggested selections of topics from across the text, which can be used to construct stand-alone courses with particular emphasis. Other options are possible based on individual preferences.

(1) **Iris dataset**. This classical dataset contains information about the sepal length and width for three types of iris flowers: virginica, setosa, and versicolor. It was originally used by Fisher (1936) and is available at the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/iris. Actually, three of the datasets in our list are available from this useful repository — see Dua and Graff (2019).

(2) **MNIST dataset**. This is a second popular dataset, which is useful for classifying handwritten digits. It was used in the work by LeCun *et al.* (1998) on document recognition. The data contains 60,000 labeled training examples and 10,000 labeled test examples for the digits 0 through 9. It can be downloaded from http://yann.lecun.com/exdb/mnist/.

(3) **CIFAR-10 dataset**. This dataset consists of color images that can belong to one of 10 classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. It is described by Krizhevsky (2009) and can be downloaded from www.cs.toronto.edu/~kriz/cifar.html.

(4) **FBI crime dataset**. This dataset contains statistics showing the burglary rates per 100,000 inhabitants during 1997–2016. The source of the data is the US Criminal Justice Information Services Division at https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-1.

(5) **Sea level and global temperature changes dataset**. The sea level dataset measures the change in sea level relative to the start of 1993. There are 952 data points consisting of fractional year values. The source of the data is the NASA Goddard Space Flight Center at https://climate.nasa.gov/vital-signs/sea-level/. For information on how the data was generated, the reader may consult Beckley *et al.* (2017) and the report GSFC (2017). The temperature dataset measures changes in the global surface temperature relative to the average over the period 1951–1980. There are 139 measurements between the years 1880 and 2018. The source of the data is the NASA Goddard Institute for Space Studies (GISS) at https://climate.nasa.gov/vital-signs/global-temperature/.

(6) **Breast cancer Wisconsin dataset**. This dataset consists of 569 samples, with each sample corresponding to a benign or malignant cancer classification. It can be downloaded from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. For information on how the data was generated, the reader may consult Mangasarian, Street, and Wolberg (1995).

(7) **Heart-disease Cleveland dataset**. The dataset consists of 297 samples that belong to patients with and without heart disease. It is available on the UCI Machine Learning Repository and can be downloaded from https://archive.ics.uci.edu/ml/datasets/heart+Disease. The investigators responsible for the collection of the data are the four leading co-authors of the article by Detrano *et al.* (1989).

P.6 ACKNOWLEDGMENTS

A project of this magnitude is not possible without the support of a web of colleagues and students. I am indebted to all of them for their input at various stages of this project, either through feedback on earlier drafts or through conversations that deepened my understanding of the topics. I am grateful to several of my former and current Ph.D. students and post-doctoral associates in no specific order: Stefan Vlaski, Kun Yuan, Bicheng Ying, Zaid Towfic, Jianshu Chen, Xiaochuan Zhao, Sulaiman Alghunaim, Qiyue Zou, Zhi Quan, Federico Cattivelli, Lucas Cassano, Roula Nassif, Virginia Bordignon, Elsa Risk, Mert Kayaalp, Hawraa Salami, Mirette Sadek, Sylvia Dominguez, Sheng-Yuan Tu, Waleed Younis, Shang-Kee Tee, Chung-Kai Tu, Alireza Tarighat, Nima Khajehnouri, Vitor Nascimento, Ricardo Merched, Cassio Lopes, Nabil Yousef, Ananth Subramanian, Augusto Santos, and Mansour Aldajani. I am also indebted to former internship and visiting undergraduate and MS students Mateja Ilic, Chao Yutong, Yigit Efe Erginbas, Zhuoyoue Wang, and Edward Nguyen for their help with some of the simulations.

I also wish to acknowledge several colleagues with whom I have had fruitful interactions over the years on topics of relevance to this text, including coauthoring joint publications, and who have contributed directly or indirectly to my work: Thomas Kailath (Stanford University, USA), Vince Poor (Princeton University, USA), José Moura (Carnegie Mellon University, USA), Mos Kaveh (University of Minnesota, USA), Bernard Widrow (Stanford University, USA), Simon Haykin (McMaster University, Canada), Thomas Cover (Stanford University, USA, *in memoriam*), Gene Golub (Stanford University, USA, *in memoriam*), Sergios Theodoridis (University of Athens, Greece), Vincenzo Matta (University of Salerno, Italy), Abdelhak Zoubir (Technical University Darmstadt, Germany), Cedric Richard (Universite Côte d'Azur, France), John Treichler (Raytheon, USA), Tiberiu Constantinescu (University of Texas Dallas, USA, *in memoriam*), Sanjit Mitra and Shiv Chandrasekaran (University of California, Santa Barbara, USA), Ming Gu (University of California, Berkeley, USA), P. P. Vaidyanathan and Babak Hassibi (Caltech, USA), Jeff Shamma (University of Illinois, Urbana Champaign, USA), Hanoch Lev-Ari (Northeastern University, USA), Markus Rupp (Tech. Unisersitaet Wien, Austria), Alan Laub, Wotao Yin, Lieven Vandenberghe, Mihaela van der Schaar, and Vwani Roychowdhury (University of California, Los Angeles), Vitor Nascimento (University of São Paulo, Brazil), Jeronimo Arena Garcia (Universidad Carlos III, Spain), Tareq Al-Naffouri (King Abdullah University of Science and Technology, Saudi Arabia), Jie Chen (Northwestern Polytechnical University, China), Sergio Barbarossa (Universita di Roma, Italy), Paolo Di Lorenzo (Universita di Roma, Italy), Alle-Jan van der Veen (Delft University, the Netherlands), Paulo Diniz (Federal University of Rio de Janeiro, Brazil), Sulyman Kozat (Bilkent University, Turkey), Mohammed Dahleh (University of California, Santa Bar-

bara, USA, *in memoriam*), Alexandre Bertrand (Katholieke Universiteit Leuven, Belgium), Marc Moonen (Katholieke Universiteit Leuven, Belgium), Phillip Regalia (National Science Foundation, USA), Martin Vetterli, Michael Unser, Pascal Frossard, Pierre Vandergheynst, Rudiger Urbanke, Emre Telatar, and Volkan Cevher (EPFL, Switzerland), Helmut Bölcskei (ETHZ, Switzerland), Visa Koivunen (Aalto University, Finland), Isao Yamada (Tokyo Institute of Technology, Japan), Zhi-Quan Luo and Shuguang Cui (Chinese University of Hong Kong, Shenzhen, China), Soummya Kar (Carnegie Mellon University, USA), Waheed Bajwa (Rutgers University, USA), Usman Khan (Tufts University, USA), Michael Rabbat (Facebook, Canada), Petar Djuric (Stony Brook University, NY), Lina Karam (Arizona State University, USA, and Lebanese American University, Lebanon), Danilo Mandic (Imperial College, United Kingdom), Jonathon Chambers (University of Leicester, United Kingdom), Rabab Ward (University of British Columbia, Canada), and Nikos Sidiropoulos (University of Virginia, USA).

I would like to acknowledge the support of my publisher, Elizabeth Horne, at Cambridge University Press during the production phase of this project. I would also like to express my gratitude to the publishers IEEE, Pearson Education, NOW, and Wiley for allowing me to adapt some excerpts and problems from my earlier works, namely, Sayed (*Fundamentals of Adaptive Filtering*, ©2003 Wiley), Sayed (*Adaptive Filters*, ©2008 Wiley), Sayed (*Adaptation, Learning, and Optimization over Networks*, ©2014 A. H. Sayed by NOW Publishers), Sayed ("Big ideas or big data," ©2018 IEEE), and Kailath, Sayed, and Hassibi (*Linear Estimation*, ©2000 Prentice Hall).

I initiated my work on this project in Westwood, Los Angeles, while working as a faculty member at the University of California, Los Angeles (UCLA), and concluded it in Lausanne, Switzerland, while working at the École Polytechnique Fédérale de Lausanne (EPFL). I am grateful to both institutions for their wonderful and supportive environments.

My wife Laila, and daughters Faten and Samiya, have always provided me with their utmost support and encouragement without which I would not have been able to devote my early mornings and good portions of my weekend days to the completion of this text. My beloved parents, now deceased, were overwhelming in their support of my education. For all the sacrifices they have endured during their lifetime, I dedicate this text to their loving memory, knowing very well that this tiny gift will never match their gift.

Ali H. Sayed
Lausanne, Switzerland
March 2021

References

Aronszajn, N. (1950), "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404.

Bayes, T. and R. Price (1763), "An essay towards solving a problem in the doctrine of chances," Bayes's article communicated by R. Price and published posthumously in *Philos. Trans. Roy. Soc. Lond.*, vol. 53, pp. 370–418.

Beckley, B. D., P. S. Callahan, D. W. Hancock, G. T. Mitchum, and R. D. Ray (2017), "On the cal-mode correction to TOPEX satellite altimetry and its effect on the global mean sea level time series," *J. Geophy. Res. Oceans*, vol. 122, no. 11, pp. 8371–8384.

Bedeian, A. G. (2016), "A note on the aphorism "there is nothing as practical as a good theory"," *J. Manag. Hist.*, pp. 236–242.

Bellman, R. E. (1957a), *Dynamic Programming*, Princeton University Press. Also published in 2003 by Dover Publications.

Bellman, R. E. (1957b), "A Markovian decision process," *Indiana Univ. Math. J.*, vol. 6, no. 4, pp. 679–684.

Cauchy, A.-L. (1847), "Methode générale pour la résolution des systems déquations simultanes," *Comptes Rendus Hebd. Séances Acad. Sci.*, vol. 25, pp. 536–538.

Cooley, J. W. and J. W. Tukey (1965), "An algorithm for the machine calculation of complex Fourier series" *Math. Comput.*, vol. 19, no. 90, pp. 297–301.

Daintith, J. (2008), editor, *Biographical Encyclopedia of Scientists*, 3rd ed., CRC Press.

de Moivre, A. (1730), *Miscellanea Analytica de Seriebus et Quadraturis*, J. Tonson and J. Watts, London.

Detrano, R., A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher (1989), "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, pp. 304–310.

Dua, D. and C. Graff (2019), *UCI Machine Learning Repository*, available at http://archive.ics.uci.edu/ml, School of Information and Computer Science, University of California, Irvine.

Einstein, A. (1916), "Näherungsweise Integration der Feldgleichungen der Gravitation," *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften Berlin*, part 1, pp. 688–696.

Fisher, R. A. (1912), "On an absolute criterion for fitting frequency curves," *Messeg. Math.*, vol. 41, pp. 155–160.

Fisher, R. A. (1922), "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. Lond. Ser. A.*, vol. 222, pp. 309–368.

Fisher, R. A. (1925), "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, vol. 22, pp. 700–725.

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188.

Fourier, J. (1822), *Théorie Analytique de la Chaleur.*, Firmin Didot Père et Fils. English translation by A. Freeman in 1878 reissued as *The Analytic Theory of Heat*, Dover Publications.

Gauss, C. F. (1903), *Carl Friedrich Gauss Werke,* Akademie der Wissenschaften.

GSFC (2017), "Global mean sea level trend from integrated multi-mission ocean altimeters TOPEX/Poseidon, Jason-1, OSTM/Jason-2," ver. 4.2 PO.DAAC, CA, USA. Dataset accessed 2019-03-18 at http://dx.doi.org/10.5067/GMSLM-TJ42.

Hall, T. (1970), *Carl Friedrich Gauss: A Biography*, MIT Press.

Heath, J. L. (1912), *The Works of Archimedes*, Dover Publications.

Kailath, T. (1968), "An innovations approach to least-squares estimation, part I: Linear filtering in additive white noise," *IEEE Trans. Aut. Control*, vol. 13, pp. 646–655.

Kailath, T. (1971), "RKHS approach to detection and estimation problems I – Deterministic signals in Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 17, no. 5, pp. 530–549.

Kailath, T., A. H. Sayed, and B. Hassibi (2000), *Linear Estimation*, Prentice Hall.

Kalman, R. E. (1960), "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, pp. 34–45.

Kolmogorov, A. N. (1939), "Sur l'interpolation et extrapolation des suites stationnaires," *C. R. Acad. Sci.*, vol. 208, p. 2043.

Krizhevsky, A. (2009), *Learning Multiple Layers of Features from Tiny Images*, MS dissertation, Computer Science Department, University of Toronto, Canada.

Laplace, P. S. (1774), "Mémoire sur la probabilité des causes par les événements," *Mém. Acad. R. Sci. de MI (Savants étrangers)*, vol. 4, pp. 621–656. See also *Oeuvres Complètes de Laplace*, vol. 8, pp. 27–65 published by the L'Académie des Sciences, Paris, during the period 1878–1912. Translated by S. M. Sitgler, *Statistical Science*, vol. 1, no. 3, pp. 366–367.

Laplace, P. S. (1812), *Théorie Analytique des Probabilités*, Paris.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998), "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324.

Levinson, N. (1947), "The Wiener RMS error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278.

Lewin, K. (1945), "The research center for group dynamics at MIT," *Sociometry*, vol. 8, pp. 126–135. See also page 169 of Lewin, K. (1952), *Field Theory in Social Science: Selected Theoretical Papers by Kurt Lewin*, Tavistock.

Lyapunov, A. M. (1901), "Nouvelle forme du théoreme sur la limite de probabilité," *Mémoires de l'Académie de Saint-Petersbourg*, vol. 12, no. 8, pp. 1–24.

Mangasarian, O. L., W. N. Street, and W. H. Wolberg (1995), "Breast cancer diagnosis and prognosis via linear programming," *Op. Res.*, vol. 43, no. 4, pp. 570–577.

Markov, A. A. (1913), "An example of statistical investigation in the text of *Eugene Onyegin* illustrating coupling of texts in chains," *Proc. Acad. Sci. St. Petersburg*, vol. 7, no. 3, p. 153–162. English translation in *Science in Context*, vol. 19, no. 4, pp. 591–600, 2006.

Mercer, J. (1909), "Functions of positive and negative type and their connection with the theory of integral equations," *Philos. Trans. Roy. Soc. Lond. Ser. A*, vol. 209, pp. 415–446.

Nyquist, H. (1928), "Certain topics in telegraph transmission theory," *Trans. AIEE*, vol. 47, pp. 617–644. Reprinted as classic paper in *Proc. IEEE*, vol. 90, no. 2, pp. 280–305, Feb. 2002.

Parzen, E. (1962), "Extraction and detection problems and reproducing kernel Hilbert spaces," *J. Soc. Indus. Appl. Math. Ser. A: Control*, vol. 1, no. 1, pp. 35–62.

Pearson, K. (1894), "Contributions to the mathematical theory of evolution," *Philos. Trans. Roy. Soc. London*, vol. 185, pp. 71–110.

Pearson, K. (1896), "Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia," *Philos. Trans. Roy. Soc. London*, vol. 187, pp. 253–318.

Pearson, K. (1901), "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572.

Porter, R. and M. Ogilvie (2000), editors, *The Biographical Dictionary of Scientists*, 3rd ed., Oxford University Press.

Robbins, H. and S. Monro (1951), "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–407.

Rosenblatt, F. (1957), *The Perceptron: A Perceiving and Recognizing Automaton*, Technical Report 85-460-1, Project PARA, Cornell Aeronautical Lab.

Sayed, A. H. (2003), *Fundamentals of Adaptive Filtering*, Wiley.

Sayed, A. H. (2008), *Adaptive Filters*, Wiley.

Sayed, A. H. (2014), *Adaptation, Learning, and Optimization over Networks,* Foundations and Trends in Machine Learning, NOW Publishers, vol. 7, no. 4–5, pp. 311–801.

Sayed, A. H. (2018), "Big ideas or big data?" *IEEE Signal Process. Mag.*, vol. 35, no. 2, pp. 5–6.

Shannon, C. E. (1948a), "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423.

Shannon, C. E. (1948b), "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656.

Viterbi, A. J. (1967), "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, pp. 260–269.

Widrow, B. and M. E. Hoff (1960), "Adaptive switching circuits," *IRE WESCON Conv. Rec.*, Institute of Radio Engineers, pt. 4, pp. 96–104.

Wiener, N. (1949), *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* Technology Press and Wiley. Originally published in 1942 as a classified Nat. Defense Res. Council Report. Also published under the title *Time Series Analysis* by MIT Press.

Zadeh, L. A. (1954), "System theory," *Columbia Engr. Quart.*, vol. 8, pp. 16–19.