Inference and Learning from Data

Volume I

This extraordinary three-volume work, written in an engaging and rigorous style by a world authority in the field, provides an accessible, comprehensive introduction to the full spectrum of mathematical and statistical techniques underpinning contemporary methods in data-driven learning and inference.

This first volume, *Foundations*, introduces core topics in inference and learning, such as matrix theory, linear algebra, random variables, convex optimization, stochastic optimization, and decentralized methods, and prepares students for studying their practical application in later volumes.

A consistent structure and pedagogy is employed throughout this volume to reinforce student understanding, with over 600 end-of-chapter problems (including solutions for instructors), 180 solved examples, 100 figures, datasets, and downloadable Matlab code. Supported by sister volumes *Inference* and *Learning*, and unique in its scale and depth, this textbook sequence is ideal for early-career researchers and graduate students across many courses in signal processing, machine learning, statistical analysis, data science, and inference.

Ali H. Sayed is Professor and Dean of Engineering at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He has also served as Distinguished Professor and Chairman of Electrical Engineering at the University of California, Los Angeles (UCLA), USA, and as President of the IEEE Signal Processing Society. He is a member of the US National Academy of Engineering (NAE) and The World Academy of Sciences (TWAS), and a recipient of several awards, including the 2022 IEEE Fourier Award and the 2020 IEEE Norbert Wiener Society Award. He is a Fellow of the IEEE, EURASIP, and AAAS.

Inference and Learning from Data

Volume I: Foundations

ALI H. SAYED
École Polytechnique Fédérale de Lausanne
University of California at Los Angeles

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009218122
DOI: 10.1017/9781009218146

First published 2023

Printed in the United Kingdom by Bell and Bain Ltd

A catalogue record for this publication is available from the British Library.

ISBN - 3 Volume Set 978-1-009-21810-8 Hardback
ISBN - Volume I 978-1-009-21812-2 Hardback
ISBN - Volume II 978-1-009-21826-9 Hardback
ISBN - Volume III 978-1-009-21828-3 Hardback

Additional resources for this publication at www.cambridge.org/sayed-vol1

In loving memory of my parents

Contents

Cambridge University Press & Assessment
978-1-009-21812-2 — Inference and Learning from Data
Volume 1
Frontmatter
More Information

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 toward 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 explosive 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 (1730), 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 (1949) 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 from 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 most 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 multilayer 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 driving 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 underpinnings 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-