

# Table of Contents

<b>Preface to the Second Edition</b> .....	xiii
<b>Preface to the First Edition</b> .....	xvii
<b>1. Introduction</b> .....	1
<b>2. Preliminaries</b> .....	13
2.1 Pseudometric Spaces, Packing and Covering Numbers .....	13
2.1.1 Pseudometric Spaces .....	13
2.1.2 Packing and Covering Numbers .....	14
2.1.3 Compact and Totally Bounded Sets .....	16
2.2 Probability Measures .....	17
2.2.1 Definition of a Probability Space .....	17
2.2.2 A Pseudometric Induced by a Probability Measure ...	18
2.2.3 A Metric on the Set of Probability Measures .....	19
2.2.4 Random Variables .....	21
2.2.5 Conditional Expectations .....	23
2.3 Large Deviation Type Inequalities .....	24
2.3.1 Chernoff Bounds .....	24
2.3.2 Chernoff-Okamoto Bound .....	26
2.3.3 Hoeffding's Inequality .....	26
2.4 Stochastic Processes, Almost Sure Convergence .....	29
2.4.1 Probability Measures on Infinite Cartesian Products ..	29
2.4.2 Stochastic Processes .....	29
2.4.3 The Borel-Cantelli Lemma and Almost Sure Conver-	
gence .....	30
2.5 Mixing Properties of Stochastic Processes .....	33
2.5.1 Definitions of Various Kinds of Mixing Coefficients ...	34
2.5.2 Inequalities for Mixing Processes .....	36
<b>3. Problem Formulations</b> .....	43
3.1 Uniform Convergence of Empirical Means .....	43
3.1.1 The UCEM Property .....	43
3.1.2 The UCEMUP Property .....	52

3.1.3	Extension to Dependent Input Sequences . . . . .	54
3.2	Learning Concepts and Functions . . . . .	55
3.2.1	Concept Learning . . . . .	55
3.2.2	Function Learning . . . . .	64
3.2.3	Extension to Dependent Input Sequences . . . . .	65
3.2.4	Assumptions Underlying the Model of Learning . . . . .	66
3.2.5	Alternate Notions of Learnability . . . . .	70
3.3	Model-Free Learning . . . . .	76
3.3.1	Problem Formulation . . . . .	76
3.3.2	Relationship to the Uniform Convergence of Empirical Means . . . . .	81
3.4	Preservation of UCEMUP and PAC Properties . . . . .	83
3.4.1	Preservation of UCEMUP Property with Beta-Mixing Inputs . . . . .	84
3.4.2	Law of Large Numbers Under Alpha-Mixing Inputs . . . . .	89
3.4.3	Preservation of PAC Learning Property with Beta-Mixing Inputs . . . . .	94
3.4.4	Preservation of PAC Learning Property with Beta-Mixing Inputs: Continued . . . . .	95
3.4.5	Replacing $\mathcal{P}$ by its Closure . . . . .	97
3.5	Markov Chains and Beta-Mixing . . . . .	100
3.5.1	Geometric Ergodicity and Beta-Mixing . . . . .	100
3.5.2	Beta-Mixing Properties of Markov Sequences . . . . .	105
3.5.3	Mixing Properties of Hidden Markov Models . . . . .	110
<b>4.</b>	<b>Vapnik-Chervonenkis, Pseudo- and Fat-Shattering Dimensions . . . . .</b>	<b>115</b>
4.1	Definitions . . . . .	115
4.1.1	The Vapnik-Chervonenkis Dimension . . . . .	115
4.1.2	The Pseudo-Dimension . . . . .	120
4.1.3	The Fat-Shattering Dimension . . . . .	122
4.2	Bounds on Growth Functions . . . . .	123
4.2.1	Growth Functions of Collections of Sets . . . . .	123
4.2.2	Bounds on Covering Numbers Based on the Pseudo-Dimension . . . . .	128
4.2.3	Metric Entropy Bounds for Families of Functions . . . . .	132
4.2.4	Bounds on Covering Numbers Based on the Fat-Shattering Dimension . . . . .	139
4.3	Growth Functions of Iterated Families . . . . .	141
<b>5.</b>	<b>Uniform Convergence of Empirical Means . . . . .</b>	<b>149</b>
5.1	Restatement of the Problems Under Study . . . . .	149
5.2	Equivalence of the UCEM and ASCEM Properties . . . . .	153
5.3	Main Theorems . . . . .	155
5.4	Preliminary Lemmas . . . . .	161

5.5	Theorem 5.1: Proof of Necessity . . . . .	173
5.6	Theorem 5.1: Proof of Sufficiency . . . . .	178
5.7	Proofs of the Remaining Theorems . . . . .	190
5.8	Uniform Convergence Properties of Iterated Families . . . . .	194
5.8.1	Boolean Operations on Collections of Sets . . . . .	195
5.8.2	Uniformly Continuous Mappings on Families of Functions . . . . .	196
5.8.3	Families of Loss Functions . . . . .	200
<b>6.</b>	<b>Learning Under a Fixed Probability Measure . . . . .</b>	<b>207</b>
6.1	Introduction . . . . .	207
6.2	UCEM Property Implies ASEC Learnability . . . . .	209
6.3	Finite Metric Entropy Implies Learnability . . . . .	216
6.4	Consistent Learnability . . . . .	224
6.4.1	Consistent PAC Learnability . . . . .	224
6.4.2	Consistent PUAC Learnability . . . . .	226
6.5	Examples . . . . .	230
6.6	Learnable Concept Classes Have Finite Metric Entropy . . . . .	236
6.7	Model-Free Learning . . . . .	242
6.7.1	A Sufficient Condition for Learnability . . . . .	244
6.7.2	A Necessary Condition . . . . .	248
6.8	Dependent Inputs . . . . .	250
6.8.1	Finite Metric Entropy and Alpha-Mixing Input Sequences . . . . .	250
6.8.2	Consistent Learnability and Beta-Mixing Input Sequences . . . . .	251
<b>7.</b>	<b>Distribution-Free Learning . . . . .</b>	<b>255</b>
7.1	Uniform Convergence of Empirical Means . . . . .	255
7.1.1	Function Classes . . . . .	256
7.1.2	Concept Classes . . . . .	258
7.1.3	Loss Functions . . . . .	261
7.2	Function Learning . . . . .	263
7.2.1	Finite P-Dimension Implies PAC and PUAC Learnability . . . . .	264
7.2.2	Finite P-Dimension is not Necessary for PAC Learnability . . . . .	267
7.3	Concept Learning . . . . .	269
7.3.1	Improved Upper Bound for the Sample Complexity . . . . .	269
7.3.2	A Universal Lower Bound for the Sample Complexity . . . . .	273
7.3.3	Learnability Implies Finite VC-Dimension . . . . .	278
7.4	Learnability of Functions with a Finite Range . . . . .	280

<b>8. Learning Under an Intermediate Family of Probabilities ..</b>	<b>285</b>
8.1 General Families of Probabilities .....	287
8.1.1 Uniform Convergence of Empirical Means .....	287
8.1.2 Function Learning .....	288
8.1.3 Concept Learning .....	292
8.2 Totally Bounded Families of Probabilities .....	297
8.3 Families of Probabilities with a Nonempty Interior .....	308
<b>9. Alternate Models of Learning .....</b>	<b>311</b>
9.1 Efficient Learning .....	312
9.1.1 Definition of Efficient Learnability .....	313
9.1.2 The Complexity of Finding a Consistent Hypothesis ..	317
9.2 Active Learning .....	326
9.2.1 Fixed-Distribution Learning .....	329
9.2.2 Distribution-Free Learning .....	332
9.3 Learning with Prior Information: Necessary and Sufficient Conditions .....	335
9.3.1 Definition of Learnability with Prior Information ....	335
9.3.2 Some Simple Sufficient Conditions .....	337
9.3.3 Dispersability of Function Classes .....	341
9.3.4 Connections Between Dispersability and Learnability WPI .....	344
9.3.5 Distribution-Free Learning with Prior Information ....	348
9.4 Learning with Prior Information: Bounds on Learning Rates .	352
<b>10. Applications to Neural Networks .....</b>	<b>365</b>
10.1 What is a Neural Network? .....	366
10.2 Learning in Neural Networks .....	369
10.2.1 Problem Formulation .....	369
10.2.2 Reprise of Sample Complexity Estimates .....	372
10.2.3 Complexity-Theoretic Limits to Learnability .....	377
10.3 Estimates of VC-Dimensions of Families of Networks .....	381
10.3.1 Multi-Layer Perceptron Networks .....	382
10.3.2 A Network with Infinite VC-Dimension .....	388
10.3.3 Neural Networks as Verifiers of Formulas .....	390
10.3.4 Neural Networks with Piecewise-Polynomial Activa- tion Functions .....	396
10.3.5 A General Approach .....	402
10.3.6 An Improved Bound .....	406
10.3.7 Networks with Pfaffian Activation Functions .....	410
10.3.8 Results Based on Order-Minimality .....	413
10.4 Structural Risk Minimization .....	415

<b>11. Applications to Control Systems</b> .....	421
11.1 Randomized Algorithms for Robustness Analysis .....	421
11.1.1 Introduction to Robust Control .....	421
11.1.2 Some NP-Hard Problems in Robust Control .....	424
11.1.3 Randomized Algorithms for Robustness Analysis .....	426
11.2 Randomized Algorithms for Robust Controller Synthesis: General Approach .....	429
11.2.1 Paradigm of Robust Controller Synthesis Problem ....	429
11.2.2 Various Types of “Near” Minima .....	432
11.2.3 A General Approach to Randomized Algorithms .....	435
11.2.4 Two Algorithms for Finding Probably Approximate Near Minima .....	436
11.3 VC-Dimension Estimates for Problems in Robust Controller Synthesis .....	438
11.3.1 A General Result .....	438
11.3.2 Robust Stabilization .....	438
11.3.3 Weighted $H_\infty$ -Norm Minimization .....	441
11.3.4 Weighted $H_2$ -Norm Minimization .....	444
11.3.5 Sample Complexity Considerations .....	445
11.3.6 Robust Controller Design Using Randomized Algorithms: An Example .....	449
11.4 A Learning Theory Approach to System Identification .....	453
11.4.1 Problem Formulation .....	453
11.4.2 A General Result .....	455
11.4.3 Sufficient Conditions for the UCEM Property .....	458
11.4.4 Bounds on the P-Dimension .....	461
<b>12. Some Open Problems</b> .....	465

## Preface to the Second Edition

In the roughly five years since the first edition of this book was published, several significant advances have taken place in statistical learning theory. New approaches have been successfully evolved, and some of the open problems stated in the first edition have been solved. In view of these developments, it has been decided to bring out a second edition of the book.

Compared to the first edition, here are some of the specific changes that have been made in the book. First, the substantial changes:

- At the time the first edition was published, practically all of statistical learning theory was based on the assumption that the samples to the learning algorithm were independent and identically distributed. Clearly the assumption that the learning samples are statistically independent is a very serious restriction, that deserved to be removed at the earliest opportunity. In the present edition, the notion of independence is replaced by the weaker notion of *mixing*, and it is shown that most of the main results of statistical learning theory continue to hold under this weaker hypothesis. Thus it becomes of interest to study whether stochastic processes of “practical” interest have this mixing property. It is shown that state sequences of Markov chains and output sequences of hidden Markov models both possess the mixing property, under appropriate conditions. Most of the relevant material is introduced in Chapters 2 and 3; however, in almost all chapters, the consequences of replacing i.i.d. input sequences by mixing input sequences are explored.
- The application of statistical learning theory to control systems was in a nascent state when the first edition was written. Since that time, there have been some major advances in this area. Two such advances are highlighted in the present edition. The first pertains to the use of “randomized” algorithms to provide probabilistic solutions to controller synthesis problems that are NP-hard in their deterministic form. If one insists on finding an algorithm that works all the time (i.e., a deterministic algorithm that is guaranteed to find a solution for every problem instance), then many simple-looking problems in robust controller synthesis are by now known to be NP-hard, and thus intractable unless  $P = NP$  (which most people don’t believe). On the other hand, if one is willing to settle for an algorithm that “works reasonably well most of the time” (i.e., a randomized algorithm),

then all of these problems become tractable in the sense that there exist efficient (polynomial-time) randomized algorithms. The second theme is studying the problem of system identification as a problem in statistical learning theory. System identification is a mature and well-established discipline, and one might wonder why a new approach is needed at all. The reason is that, by tradition, system identification theory is addressed to the derivation of *asymptotic* results, that tell us what happens in the limit, i.e., as the number of samples approaches infinity. However, if one is interested in combining system identification with robust control, then it is essential to have *finite-time* estimates, of the kind provided by statistical learning theory. Thus by recasting the problem of identifying an unknown system as a learning problem, it becomes possible to derive finite-time estimates for the rate at which the identified model converges to the unknown system that is being identified. These estimates can then be used to design robust controllers.

Now for more minor changes:

- The concept of “fat-shattering dimension” is introduced, and its application to learning real-valued functions is explained.
- In Chapter 9, the section on learning with prior information has been expanded to include recent results, which provide both necessary and sufficient conditions for learning with prior information.
- The chapter on neural networks has been modified to reflect recent advances.
- Since several of the open problems stated in Chapter 12 have since been solved, this chapter is thoroughly revamped.

As a result of all these changes, the pedagogical approach of the book has been substantially altered from the first edition. In that work, I attempted to build up the level gradually, whereby the first two chapters (after the Introduction) could be skipped by experts, who could proceed directly to the later chapters. However, in the present edition advanced material can be found in *every* chapter. I had to choose between introducing each new result in a place that appeared most natural to me, and retaining the monotonicity of the “difficulty function.” I opted for the former approach.

In view of the length of the book, I made a conscious decision to leave out a discussion of support vector machines, which represent an important advance in machine learning. The interested reader is referred to [46, 172, 49].

Since the publication of the first edition, I have been fortunate to have initiated a continuing collaboration with Rajeeva Karandikar of the Indian Statistical Institute, Delhi. I would like to thank him for furthering my education in probability theory, and for replacing my “seat of the pants” approach to probability theory with something more rigorous. I would also like to thank him for reading Chapters 2 and 3, and also for permitting the inclusion of sev-

eral previously unpublished results in these chapters. Any remaining errors are of course my own responsibility.

As always I would like to thank my wife Shakunthala for her consistent support throughout my career. I would also like to thank my employers, Tata Consultancy Services, especially my CEO, Mr. S. Ramadorai, for having the vision to encourage an activity of this kind in a software company.

I take this opportunity to dedicate this book to my parents, who have made me what I am. My father, Professor M. V. Subbarao, continues to be an active researcher in number theory even at the age of eighty. From his example I learnt to aspire to a career in research when I was still a boy. My mother, Mrs. Suseela Subbarao, not only played a major role in my precocity, but also passed on to me a wonderful religious lineage, whose value cannot be measured by any worldly yardstick.

Hyderabad, India  
March 2002



# Preface to the First Edition

The objective of this book is to present a comprehensive treatment of some recent developments in statistical learning theory, and their applications to analyzing the ability of neural networks to generalize; in addition, some potential applications to control systems are also sketched, and some problems for future research are indicated. The book is aimed at engineers, computer scientists, and applied mathematicians who have an interest in the broad area of machine learning. The background required to read and understand this book consists primarily of a basic understanding of the elements of probability theory. It should be emphasized that, while learning theory as discussed here uses the *formalism* of probability theory, most of the deep concepts in probability are *not* used. Chapter 2 gives a summary of the background required of the reader. The book can either be used for self-study or as a text in an advanced graduate course. After the first four chapters, the remaining chapters are more or less independent of each other, so that a reader or instructor will be able to pick and choose according to their requirements.

It might be desirable to summarize how the subject of learning theory came to its current status. Even as the modern digital computer was being invented, the scientific community was beginning its attempts to formulate mathematical theories of how machines can be made to “learn” and to “generalize” on the basis of past experience. As early as 1943, J. C. McCulloch and W. Pitts [130] studied interconnections of switching elements that were simple approximations of the biological neurons found in the human brain, and proved that every finite-state automaton can be approximated by such a network, and vice versa. In the late 1950’s Frank Rosenblatt introduced the perceptron and proved that, under suitable circumstances, a perceptron could learn to separate positive examples from negative examples. Very shortly thereafter, the training of perceptrons was given a statistical flavour by the Russian school, in a manner strongly reminiscent of subsequent developments more than twenty years later. The subject of “inductive” learning was sought to be formulated as a counterpart to deductive learning on the basis of mathematical logic. While the study of the original perceptron went into a hiatus following the publication of the book *Perceptrons* by Minsky and Papert [137], other models of learning systems such as learning automata continued to be invented and studied. The subject of neural networks, which can be

thought of as an intellectual successor to perceptron theory, was revived in spectacular fashion following the publication of the multi-volume book *Parallel Distributed Processing* by Rumelhart and McClelland [168], [169]. At least part of the appeal of neural networks stems from their claimed ability to “generalize.” A great deal of episodic evidence has been presented in the literature to support the claim that, once a neural network has been “trained” on a sufficient number of samples, it can then produce the correct output to a new, and previously unseen, input. It should be noted that, without the ability to generalize, much of the case for using neural networks would collapse – a simple table-lookup scheme would suffice if one were interested merely in constructing a network that could reproduce known input-output pairs. However, until recently there has not been a clear mathematical enunciation of just what “generalization” means, nor has there been any mathematical justification to back up the episodic evidence that neural networks seem to be able to generalize in specific situations.

While these developments were taking place in the neural networks community, the theoretical computer science community had its attention drawn to a novel formulation of the learning problem by the publication in 1984 of the paper “A Theory of the Learnable” by Leslie G. Valiant. This approach to learning has over the years come to be known as “probably approximately correct (PAC)” learning theory. In this paper, Valiant showed that Boolean functions in  $n$  variables are “learnable” in a very precise sense provided they can be expressed as a 3-CNF, that is, if they can be expressed as a conjunction of several clauses, each of which is a disjunction of no more than three variables. Several other classes of Boolean formulae were also shown to be learnable in the same sense. Since the publication of Valiant’s paper, many others have pursued the PAC formulation, refined and redefined it, derived necessary and sufficient conditions for learnability in Valiant’s and related frameworks, developed several applications, and so on.

A parallel development in the theory of empirical processes was to have a profound impact on learning theory. More than two centuries ago, J. Bernoulli showed that, if a two-sided coin is tossed repeatedly, then the fraction of “heads” converges almost surely to the *true* probability of getting “heads” as the number of tosses approaches infinity. In more modern terminology and notation, the Glivenko-Cantelli lemma of the 1930’s showed that, if one draws a sequence of random real numbers  $x_1, \dots, x_m$  in accordance with an unknown probability measure  $P$ , then the empirical distribution function converges uniformly and almost surely to the true distribution function. This result was subsequently generalized by Kolmogorov and Smirnov to vector-valued processes. Still more general problems were studied by various researchers, and culminated in the landmark 1971 paper “On the Uniform Convergence of Relative Frequencies to Their Probabilities” by V.N. Vapnik and A.Ya. Chervonenkis [194]. This paper gave necessary and sufficient conditions for the empirical estimates of the probability measures of a family of sets to con-

verge to their true values, as the number of samples approaches infinity. A combinatorial parameter, which has since come to be known as the Vapnik-Chervonenkis (VC-) dimension, plays a central role in these necessary and sufficient conditions.

The publication in 1989 of the paper “Learnability and the Vapnik-Chervonenkis Dimension” by Anselm Blumer *et al.* [32] represented another milestone in the development of PAC learning theory. This paper was apparently the first to make a connection between PAC learning theory and the theory of empirical processes. While there have been a few other papers that followed up this connection, my opinion is that by and large this connection remains unexplored, or perhaps merely unexplained to a wide audience. In particular, by reformulating the learning problem as a convergence problem for stochastic processes, it is possible to make a distinction between *information-theoretic* limitations to learning, and *complexity-theoretic* limitations to learning. Roughly speaking, information-based learning theory attempts to study what is learnable *in principle*, whereas complexity-based learning theory attempts to study what is learnable *in practice*.

By now it was widely appreciated that the PAC learning formulation presents a mathematically rigorous, as well as tractable, formulation of the intuitive idea that “neural networks can generalize.” Moreover, by estimating the VC-dimension of a neural network architecture, it is possible to make *quantitatively precise*, albeit quite conservative, estimates of the “rates” at which a neural network can “learn” and “generalize.” This naturally led several researchers to investigate ways of estimating the VC-dimension of various types of neural network architectures. The outcome of these researches is a rich theory, ranging from simple counting arguments to extremely sophisticated methods involving model theory of real numbers, algebraic geometry, and so on. It is therefore clear that the problem of estimating the VC-dimension is by now an important specialization in itself.

The issues of intelligence, learning, and generalization are also relevant in the context of control systems. In its broadest sense, a control system can be any object, natural or man-made, whose behaviour one wishes to modify into something more desirable. There are at least a few problems in control theory that can be viewed from the learning theory perspective. However, in contrast with neural networks, it is not yet clear whether the learning theory perspective offers any advantages over existing methods of control theory. Moreover, in contrast with neural networks where the “standard” PAC learning problem formulation nicely captures the notion of generalization, the problems in control theory may perhaps require a reformulation of the basic PAC learning problem.

The present monograph attempts to achieve several related objectives: (i) To present a treatment of (PAC) learning theory that brings out clearly the connection between this theory and some of the fundamental results in the theory of empirical processes; (ii) to present an application of learning theory

to generalization by neural networks; (iii) to indicate how some problems in control theory might be viewed as problems in learning, and how learning theory needs to be modified in order to be applicable to such problems; and (iv) to discuss some open problems in statistical learning theory that merit the attention of the research community. The various objectives are rather disparate in nature. In the first case, the theory of empirical processes is by now a mature subject, and an exposition of the principal results as given here will have some lasting reference value. In the second case, many new discoveries continue to be made on the computation of the VC-dimension of neural networks, and it is possible that some of the results given here will be subsumed by newer developments. Nevertheless, some of the fundamental discoveries will stand the test of time. In the third and fourth cases, the aim is more to trigger further activity than anything else.

At present there are several excellent texts on *computational* learning theory, such as [147], [9], and [99]. A deliberate choice has been made here to focus on the *statistical* aspects of learning theory, though the computational aspects are touched upon in Chapters 9 and 10. Thus the present work is intended to complement the books listed above. It should be mentioned that there is also a great deal of work in the probability theory community on the problem of nonparametric density estimation, which is closely related to the problems studied here. This body of research is not discussed at all in the present book; the interested reader is referred to [55] and the references therein. Though for the most part the book is a compendium of known results, in many places the known results are refined and/or the proofs streamlined. Thus it is hoped that both the novice as well as the expert will “learn” something from reading this book.

Now it is my pleasure to express my sincere gratitude to several individuals who have assisted me in the writing of this book. Specifically, I would like to thank (in chronological order)

- Ravi Kannan and Sanjoy Mitter for first introducing me to the fascinating world of PAC learning theory, and for encouraging me along once I got into the subject.
- Sanjeev Kulkarni for initially teaching me everything I knew about learning theory, and also for carefully reading various drafts of the book.
- Eduardo Sontag for his infectious enthusiasm for the idea of such a book, which encouraged me to complete the project in record time by my standards, and also for his careful critique of the chapter on neural networks.
- Vivek Borkar for educating me in probability theory and for serving as my own personal “oracle” (which, unlike those studied in the book, never made a mistake!).
- Dan Ocone for class-testing an early draft of the book, and giving me valuable feedback.
- Lennart Ljung and Roberto Tempo for giving me an opportunity to present this material in condensed form at their respective institutions.

- Vijay Chandru, Girish Deodhare, Piyush Gupta and S.H. Srinivasan for enthusiastically participating in the lectures given by me based on this book.
- Vishwambhar Pati for aiding me to understand the material on algebraic topology in the chapter on neural networks.

In addition, I would like to thank Wolfgang Maass for sharing several of his papers on neural networks and for useful electronic discussions, and Pascal Koiran for useful comments on the chapter on neural networks.

Finally, as for my family, what can I say in respect of their encouragement and moral support that I have not already said on several previous occasions? Sometimes I feel that my attitude towards my family mirrors that of the male protagonist in the O’Henry short story “The Pendulum.” Once again I can only say “Thank you” – it is my pleasure to dedicate this book to them.



<http://www.springer.com/978-1-85233-373-7>

Learning and Generalisation  
With Applications to Neural Networks

Vidyasagar, M.

2003, XXI, 488 p., Hardcover

ISBN: 978-1-85233-373-7