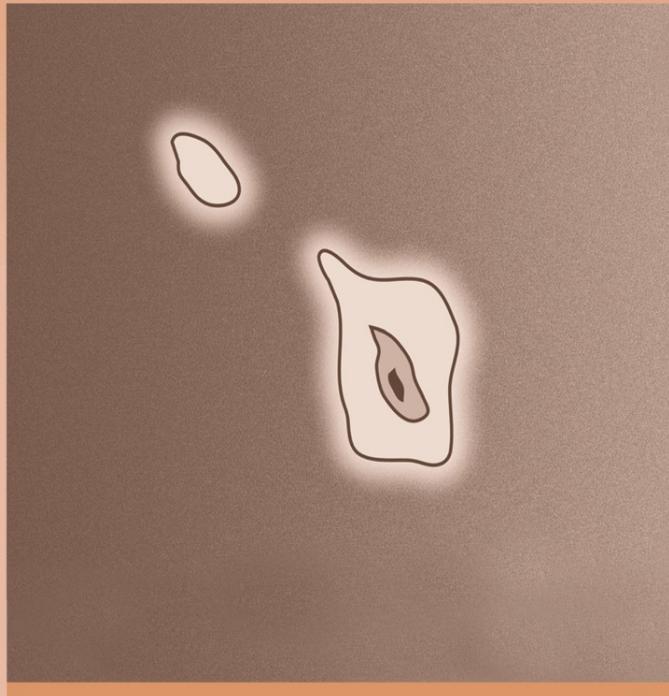


Algorithmic Learning in a Random World



Vladimir Vovk
Alex Gammerman
Glenn Shafer

Algorithmic Learning in a Random World

Algorithmic Learning in a Random World

Vladimir Vovk

*University of London
Egham, United Kingdom*

Alexander Gammerman

*University of London
Egham, United Kingdom*

Glenn Shafer

*Rutgers Business School
Newark, NJ, USA*



Springer

Vladimir Vovk
Computer Learning Research
Centre,
Dept. of Computer Science
Royal Holloway
University of London
Egham, Surrey TW2 0EX
UK

EMAIL: vovk@cs.rhul.ac.uk

Alexander Gammerman
Computer Learning Research
Centre,
Dept. of Computer Science
Royal Holloway
University of London
Egham, Surrey TW2 0EX
UK

EMAIL: alex@cs.rhul.ac.uk

Glenn Shafer
Dept. of Accounting and
Information Systems
Rutgers Business School,
Newark and New Brunswick
180 University Avenue
Newark NJ 07102

EMAIL:
gshafer@andromeda.rutgers.edu

Library of Congress Cataloging-in-Publication Data

Vovk, Vladimir.

Algorithmic Learning in a Random World / by Vladimir Vovk, Alexander Gammerman, and Glenn Shafer.

p.cm.

Includes Bibliographical references and index.

ISBN 0-387-00152-2 (HC)
ISBN-13: 978-0387-00152-4 (HC)

e-ISBN 0-387-25061-1 (eBK) Printed on acid-free paper.
e-ISBN-13: 978-038-725061-8 (eBK)

1. Prediction theory. 2. Algorithms. 3. Stochastic processes. I. Gammerman, A. (Alexander) II. Shafer, Glenn, 1946- III. Title.

QA279.2.V68 2005
519.2'87—dc22

2005042556

© 2005 Springer Science+Business Media, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America. (BS/DH)

9 8 7 6 5 4 3 2 1

SPIN 10900581 (HC) / 11399810 (eBK)

springeronline.com

Contents

Preface	XIII
List of principal results	XV
1 Introduction	1
1.1 Machine learning	1
Learning under randomness	2
Learning under unconstrained randomness	3
1.2 A shortcoming of the existing theory	3
The hold-out estimate of confidence	4
The contribution of this book	5
1.3 The on-line transductive framework	5
On-line learning	5
Transduction	6
On-line/off-line and transduction/induction compromises ..	7
1.4 Conformal prediction	7
Nested prediction sets	8
Validity	9
Efficiency	9
Conditionality	11
Flexibility of conformal predictors	11
1.5 Probabilistic prediction under unconstrained randomness ..	12
Universally consistent probabilistic predictor	12
Probabilistic prediction using a finite data set	12
Venn prediction	13
1.6 Beyond randomness	13
Testing randomness	13
Low-dimensional dynamic models	14
Islands of randomness	14
On-line compression models	15
1.7 Bibliographical remarks	15

2 Conformal prediction	17
2.1 Confidence predictors	17
Assumptions	17
Simple predictors and confidence predictors	18
Validity	19
Randomized confidence predictors	22
2.2 Conformal predictors	23
Bags	23
Nonconformity and conformity	23
p-values	25
Definition of conformal predictors	25
Validity	26
Smoothed conformal predictors	27
A general scheme for defining nonconformity	28
2.3 Ridge regression confidence machine	29
Least squares and ridge regression	29
Basic RRCM	30
Two modifications	34
Dual form ridge regression	35
Nearest neighbors regression	38
Experimental results	39
2.4 Are there other ways to achieve validity?	42
2.5 Conformal transducers	44
Normalized confidence predictors and confidence transducers	46
2.6 Proofs	47
Proof of Theorem 2.1	47
Proof of Theorem 2.6	48
Proof of Proposition 2.7	48
2.7 Bibliographical and historical remarks	49
Conformal prediction	49
Least squares and ridge regression	50
Kernel methods	50
3 Classification with conformal predictors	53
3.1 More ways of computing nonconformity scores	54
Nonconformity scores from nearest neighbors	54
Nonconformity scores from support vector machines	56
Reducing classification problems to the binary case	58
3.2 Universal predictor	59
3.3 Construction of a universal predictor	61
Preliminaries	61
Conformal prediction in the current context	62
Universal predictor	63
3.4 Fine details of confidence prediction	64

Bayes confidence predictor	69
3.5 Proofs	70
Proof of Proposition 3.2	74
Proof of Proposition 3.3	77
Proof sketch of Proposition 3.4	80
Proof sketch of Proposition 3.5	83
3.6 Bibliographical remarks	95
Examples of nonconformity measures	95
Universal predictor	95
Alternative protocols	96
Confidence and credibility	96
4 Modifications of conformal predictors	97
4.1 Inductive conformal predictors	98
The general scheme for defining nonconformity	99
4.2 Further ways of computing nonconformity scores	100
De-Bayesing	102
Bootstrap	103
Decision trees	104
Boosting	104
Neural networks	105
Logistic regression	106
4.3 Weak teachers	106
Imperfectly taught predictors	107
Weak validity	108
Strong validity	109
Iterated logarithm validity	109
Efficiency	109
4.4 Off-line conformal predictors and semi-off-line ICPs	110
4.5 Mondrian conformal predictors	114
Mondrian conformal transducers	114
Using Mondrian conformal transducers for prediction	115
Generality of Mondrian taxonomies	116
Conformal transducers	117
Inductive conformal transducers	118
Label-conditional Mondrian conformal transducers	120
Attribute-conditional Mondrian conformal transducers	120
Slow teacher	122
4.6 Proofs	123
Proof of Theorem 4.2, I: $n_k/n_{k-1} \rightarrow 1$ is sufficient	123
Proof of Theorem 4.2, II: $n_k/n_{k-1} \rightarrow 1$ is necessary	126
Proof of Theorem 4.4	127
Proof of Theorem 4.8	128
4.7 Bibliographical remarks	129
Computationally efficient hedged prediction	129

VIII Contents

Specific learning algorithms and nonconformity measures	130
Weak teachers	130
Mondrian conformal predictors	130
5 Probabilistic prediction I: impossibility results	131
5.1 Diverse data sets	132
5.2 Impossibility of estimation of probabilities	132
Binary case	133
Multi-label case	134
5.3 Proof of Theorem 5.2	135
Probability estimators and statistical tests	135
Complete statistical tests	136
Restatement of the theorem in terms of statistical tests . . .	136
The proof	138
5.4 Bibliographical remarks and addenda	138
Density estimation, regression estimation, and regression	
with deterministic objects	138
Universal probabilistic predictors	139
Algorithmic randomness perspective	140
6 Probabilistic prediction II: Venn predictors	143
6.1 On-line probabilistic prediction	145
The on-line protocol for probabilistic prediction	146
An informal look at testing calibration	147
Testing using events of small probability	148
Calibration events	150
Testing using nonnegative supermartingales	150
Predictor has no satisfactory strategy	154
6.2 On-line multiprobability prediction	156
The on-line protocol	157
Validity	158
6.3 Venn predictors	158
The problem of the reference class	159
Empirical results	161
Probabilities vs. p-values	162
6.4 A universal Venn predictor	163
6.5 Proofs	163
Proof of Theorem 6.5	163
Equivalence of the two definitions of upper probability . . .	164
Proof of Theorem 6.7	166
6.6 Bibliographical remarks	168
Testing	168
Frequentist probability	168

7	Beyond exchangeability	169
7.1	Testing exchangeability	169
	Exchangeability supermartingales	170
	Power supermartingales and the simple mixture	171
	Tracking the best power martingale	173
7.2	Low-dimensional dynamic models	177
7.3	Islands of randomness	182
	A sufficient condition for asymptotic validity	183
	Markov sequences	184
7.4	Proof of Proposition 7.2	185
7.5	Bibliographical remarks	187
8	On-line compression modeling I: conformal prediction	189
8.1	On-line compression models	190
8.2	Conformal transducers and validity of OCM	192
	Finite-horizon result	193
8.3	Repetitive structures	195
8.4	Exchangeability model and its modifications	196
	Exchangeability model	196
	Generality and specificity	197
	Inductive-exchangeability models	197
	Mondrian-exchangeability models	198
8.5	Gaussian model	199
	Gauss linear model	201
	Student predictor vs. ridge regression confidence machine .	203
8.6	Markov model	205
8.7	Proof of Theorem 8.2	211
8.8	Bibliographical remarks and addenda	214
	Kolmogorov's program	214
	Repetitive structures	215
	Exchangeability model	217
	Gaussian model	217
	Markov model	217
	Kolmogorov's modeling vs. standard statistical modeling .	218
9	On-line compression modeling II: Venn prediction	223
9.1	Venn prediction in on-line compression models	224
9.2	Generality of finitary repetitive structures	224
9.3	Hypergraphical models	225
	Hypergraphical repetitive structures	226
	Fully conditional Venn predictor	227
	Generality of hypergraphical models	227
9.4	Junction-tree models	228
	Combinatorics of junction-tree models	229
	Shuffling data sets	230

	Decomposability in junction-tree models	231
	Prediction in junction-tree models	231
	Universality of the fully conditional Venn predictor	233
9.5	Causal networks and a simple experiment	233
9.6	Proofs and further information	236
	Proof of Theorem 9.1	236
	Maximum-likelihood estimation in junction-tree models	238
9.7	Bibliographical remarks	240
	Additive models	240
10	Perspectives and contrasts	241
10.1	Inductive learning	242
	Jacob Bernoulli's learning problem	243
	Statistical learning theory	246
	The quest for data-dependent bounds	249
	The hold-out estimate	250
	On-line inductive learning	251
10.2	Transductive learning	253
	Student and Fisher	254
	Tolerance regions	256
	Transduction in statistical learning theory	258
	PAC transduction	260
	Why on-line transduction makes sense	262
10.3	Bayesian learning	264
	Bayesian ridge regression	264
	Experimental results	267
10.4	Proofs	270
	Proof of Proposition 10.1	270
	Proof of Proposition 10.2	271
10.5	Bibliographical remarks	272
	Inductive prediction	272
	Transductive prediction	273
	Bayesian prediction	273
	Appendix A: Probability theory	275
A.1	Basics	275
	Kolmogorov's axioms	275
	Convergence	277
A.2	Independence and products	277
	Products of probability spaces	277
	Randomness model	278
A.3	Expectations and conditional expectations	279
A.4	Markov kernels and regular conditional distributions	280
	Regular conditional distributions	281
A.5	Exchangeability	282

Conditional probabilities given a bag	283
A.6 Theory of martingales	284
Limit theorems	286
A.7 Hoeffding's inequality and McDiarmid's theorem	287
A.8 Bibliographical remarks	289
Conditional probabilities	289
Martingales	290
Hoeffding's inequality and McDiarmid's theorem	290
Appendix B: Data sets	291
B.1 USPS data set	291
B.2 Boston Housing data set	292
B.3 Normalization	292
B.4 Randomization and reshuffling	293
B.5 Bibliographical remarks	294
Appendix C: FAQ	295
C.1 Unusual features of conformal prediction	295
C.2 Conformal prediction vs. standard methods	296
Notation	299
References	303
Index	317

Preface

This book is about prediction algorithms that learn. The predictions these algorithms make are often imperfect, but they improve over time, and they are *hedged*: they incorporate a valid indication of their own accuracy and reliability. In most of the book we make the standard assumption of randomness: the examples the algorithm sees are drawn from some probability distribution, independently of one another. The main novelty of the book is that our algorithms learn and predict simultaneously, continually improving their performance as they make each new prediction and find out how accurate it is. It might seem surprising that this should be novel, but most existing algorithms for hedged prediction first learn from a training data set and then predict without ever learning again. The few algorithms that do learn and predict simultaneously do not produce hedged predictions. In later chapters we relax the assumption of randomness to the assumption that the data come from an on-line compression model. We have written the book for researchers in and users of the theory of prediction under randomness, but it may also be useful to those in other disciplines who are interested in prediction under uncertainty.

This book has its roots in a series of discussions at Royal Holloway, University of London, in the summer of 1996, involving AG, Vladimir N. Vapnik and VV. Vapnik, who was then based at AT&T Laboratories in New Jersey, was visiting the Department of Computer Science at Royal Holloway for a couple of months as a part-time professor. VV had just joined the department, after a year at the Center for Advanced Study in Behavioral Sciences at Stanford. AG had become the head of department in 1995 and invited both Vapnik and VV to join the department as part of his program (enthusiastically supported by Norman Gowar, the college principal) of creating a machine learning center at Royal Holloway. The discussions were mainly concerned with Vapnik's work on support vector machines, and it was then that it was realized that the number of support vectors used by such a machine could serve as a basis for hedged prediction.

Our subsequent work on this idea involved several doctoral students at Royal Holloway. Ilia Nouretdinov has made several valuable theoretical contributions. Our other students working on this topic included Craig Saunders, Tom Melluish, Kostas Proedrou, Harris Papadopoulos, David Surkov, Leo Gordon, Tony Bellotti, Daniil Ryabko, and David Lindsay. The contribution of Yura Kalnishkan and Misha Vyugin to this book was less direct, mainly through their work on predictive complexity, but still important. Thank you all.

GS joined the project only in the autumn of 2003, although he had earlier helped develop some of its key ideas through his work with VV on game-theoretic probability; see their joint book – *Probability and Finance: It's Only a Game!* – published in 2001.

Steffen Lauritzen introduced both GS and VV to repetitive structures. In VV's case, the occasion was a pleasant symposium organized and hosted by Lauritzen in Aalborg in June 1994. We have also had helpful conversations with Masafumi Akahira, Satoshi Aoki, Peter Bramley, John Campbell, Alexey Chervonenkis, Philip Dawid, José González, Thore Graepel, Gregory Gutin, David Hand, Fumiyasu Komaki, Leonid Levin, Xiao Hui Liu, George Loizou, Zhiyuan Luo, Per Martin-Löf, Sally McClean, Boris Mirkin, Fionn Murtagh, John Shawe-Taylor, Sasha Shen', Akimichi Takemura, Kei Takeuchi, Roger Thatcher, Vladimir V'yugin, David Waltz, and Chris Watkins.

Many ideas in this book have their origin in our attempts to understand the mathematical and philosophical legacy of Andrei Nikolaevich Kolmogorov. Kolmogorov's algorithmic notions of complexity and randomness (especially as developed by Martin-Löf and Levin) have been for us the main source of intuition, although they almost disappeared in the final version. VV is grateful to Andrei Nikolaevich for steering him in the direction of compression modeling and for his insistence on its independent value.

We thank the following bodies for generous financial support: EPSRC through grants GR/L35812, GR/M14937, GR/M16856, and GR/R46670; BBSRC through grant 111/BIO14428; MRC through grant S505/65; the Royal Society; the European Commission through grant IST-1999-10226; NSF through grant 5-26830.

University of London (VV, AG, and GS)
 Rutgers University (GS)
 July 2004

Vladimir Vovk
 Alexander Gammerman
 Glenn Shafer

List of principal results

<i>Theorem</i>	<i>Topic and page</i>
8.1	Conformal predictors are valid, 193
9.1	Venn predictors are valid, 224
2.6	Conformal predictors are the only valid confidence predictors in a natural class, 43
3.1	There exists an asymptotically optimal confidence predictor (an explicitly constructed conformal predictor), 61
6.7	There exists an asymptotically efficient Venn predictor (explicitly constructed in the proof), 163
4.2	Characterization of teaching schedules under which smoothed conformal predictors are asymptotically valid in probability, 108
4.4	Asymptotic validity of smoothed conformal predictors taught by weak teachers, 109
4.8	Asymptotic efficiency of smoothed conformal predictors taught by weak teachers, 110
5.2	Impossibility of probability estimation from diverse data sets under unconstrained randomness, 134
2.1	There are no exactly valid confidence predictors, 21
6.5	No probabilistic predictor is well calibrated under unconstrained randomness, 155

Introduction

In this introductory chapter, we sketch the existing work in machine learning on which we build and then outline the contents of the book.

1.1 Machine learning

The rapid development of computer technology during the last several decades has made it possible to solve ever more difficult problems in a wide variety of fields. The development of software has been essential to this progress. The painstaking programming in machine code or assembly language that was once required to solve even simple problems has been replaced by programming in high-level object-oriented languages. We are concerned with the next natural step in this progression – the development of programs that can *learn*, i.e., automatically improve their performance with experience.

The need for programs that can learn was already recognized by Alan Turing (1950), who argued that it may be too ambitious to write from scratch programs for tasks that even humans must learn to perform. Consider, for example, the problem of recognizing hand-written digits. We are not born able to perform this task, but we learn to do it quite robustly. Even when the hand-written digit is represented as a gray-scale matrix, as in Fig. 1.1, we can recognize it easily, and our ability to do so scarcely diminishes when it is slightly rotated or otherwise perturbed. We do not know how to write instructions for a computer that will produce equally robust performance.

The essential difference between a program that implements instructions for a particular task and a program that learns is adaptability. A single learn-



Fig. 1.1. A hand-written digit

ing program may be able to learn a wide variety of tasks: recognizing hand-written digits and faces, diagnosing patients in a hospital, estimating house prices, etc.

Recognition, diagnosis, and estimation can all be thought of as special cases of prediction. A person or a computer is given certain information and asked to predict the answer to a question. A broad discussion of learning would go beyond prediction to consider the problems faced by a robot, who needs to act as well as predict. The literature on machine learning, has emphasized prediction, however, and the research reported in this book is in that tradition. We are interested in algorithms that learn to predict well.

Learning under randomness

One learns from experience. This is as true for a computer as it is for a human being. In order for there to be something to learn there must be some stability in the environment; it must be governed by constant, or evolving only slowly, laws. And when we learn to predict well, we may claim to have learned something about that environment.

The traditional way of making the idea of a stable environment precise is to assume that it generates a sequence of examples randomly from some fixed probability distribution, say Q , on a fixed space of possible examples, say \mathbf{Z} . These mathematical objects, \mathbf{Z} and Q , describe the environment.

The environment can be very complex; \mathbf{Z} can be large and structured in a complex way. This is illustrated by the USPS data set from which Fig. 1.1 is drawn (see Appendix B). Here an example is any 16×16 image with 31 shades of gray, together with the digit the image represents (an integer between 0 to 9). So there are $31^{16 \times 16} \times 10$ (this is approximately 10^{382}) possible examples in the space \mathbf{Z} .

In most of this book, we assume that each example consists of an *object* and its *label*. In the USPS dataset, for example, an object is a gray-scale matrix like the one in Fig. 1.1, and its label is the integer between 0 and 9 represented by the gray-scale matrix.

In the problem of recognizing hand-written digits and other typical machine-learning problems, it is the space of objects, the space of possible gray-scale images, that is large. The space of labels is either a small finite set (in what is called *classification problems*) or the set of real numbers (*regression problems*).

When we say that the examples are chosen randomly from Q , we mean that they are independent in the sense of probability theory and all have the distribution Q . They are independent and identically distributed. We call this the *randomness assumption*.

Of course, not all work in machine learning is concerned with learning under randomness. In learning with expert advice, for example, randomness is replaced by a game-theoretic set-up (Vovk 2001a); here a typical result is that the learner can predict almost as well as the best strategy in a pool of

possible strategies. In reinforcement learning, which is concerned with rational decision-making in a dynamic environment (Sutton and Barto 1998), the standard assumption is Markovian. In this book, we will consider extensions of learning under randomness in Chaps. 7–9.

Learning under unconstrained randomness

Sometimes we make the randomness assumption without assuming anything more about the environment: we know the space of examples Z , we know that examples are drawn independently from the same distribution, and this is all we know. We know nothing at the outset about the probability distribution Q from which each example is drawn. In this case, we say we are *learning under unconstrained randomness*. Most of the work in this book, like much other work in machine learning, is concerned with learning under unconstrained randomness.

The strength of modern machine-learning methods often lies in their ability to make hedged predictions under unconstrained randomness in a *high-dimensional* environment, where examples have a very large (or infinite) number of components. We already mentioned the USPS data set, where each example consists of 257 components (16×16 pixels and the label). In machine learning, this number is now considered small, and the problem of learning from the USPS dataset is sometimes regarded as a toy problem.

1.2 A shortcoming of the existing theory

Machine learning has made significant strides in its study of learning under unconstrained randomness. We now have a wide range of algorithms that often work very well in practice: decision trees, neural networks, nearest neighbors algorithms, and naive Bayes methods have been used for decades; newer algorithms include support vector machines and boosting, an algorithm that is used to improve the quality of other algorithms.

From a theoretical point of view, machine learning's most significant contributions to learning under unconstrained randomness are comprised by *statistical learning theory*. This theory, which began with the discovery of VC dimension by Vapnik and Chervonenkis in the late 1960s and was partially rediscovered independently by Valiant (1984), has produced both deep mathematical results and learning algorithms that work very well in practice (see Vapnik 1998 for a recent review).

Given a “training” set of examples, statistical learning theory produces what we call a *prediction rule* – a function mapping the objects into the labels. Formally, the value taken by a prediction rule on a new object is a *simple prediction* – a guess that is not accompanied by any statement concerning how accurate it is likely to be. The theory does guarantee, however, that as the training set becomes bigger and bigger these predictions will become more

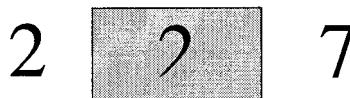


Fig. 1.2. In the problem of digit recognition, we would like to attach lower confidence to the prediction for the image in the middle than to the predictions for the images on the left and the right

and more accurate with greater and greater probability: they are *probably approximately correct*.

How probably and how approximately? This question has not been answered as well as we might like. This is because the theoretical results that might be thought to answer it, the bounds that demonstrate arbitrarily good accuracy with sufficiently large sizes of the training set, are usually too loose to tell us anything interesting for training sets that we actually have. This happens in spite of the empirical fact that the predictions often perform very well in practice. Consider, for example, the problem of recognizing hand-written digits, which we have already discussed. Here we are interested in giving an upper bound on the probability that our learning algorithm fails to choose the right digit; we might like this probability to be less than 0.05, for example, so that we can be 95% confident that the prediction is correct. Unfortunately, typical upper bounds on the probability of error provided by the theory, even for relatively clean data sets such as the USPS data set we have discussed, are greater than 1; bounds less than 1 can usually be achieved only for very straightforward problems or with very large data sets. This is true even for newer results in which the bound on the accuracy depends on the training set (as in, e.g., Littlestone and Warmuth 1986, Floyd and Warmuth 1995; cf. §10.1).

The hold-out estimate of confidence

Fortunately, there are less theoretical and more effective ways of estimating the confidence we should have in predictions output by machine-learning algorithms, including those output by the algorithms proposed by statistical learning theory. One of the most effective is the oldest and most naive: the “hold-out” estimate. In order to compute this estimate, we split the available examples into two parts, a training set and a “test” set. We apply the algorithm to the training set in order to find a prediction rule, and then we apply this prediction rule to the test set. The observed rate of errors on the test set tells us how confident we should be in the prediction rule when we apply it to new examples (for details, see §10.1).

The contribution of this book

When we use a hold-out sample to obtain a meaningful bound on the probability of error, or when we use an error bound from statistical learning theory, we are *hedging* the prediction – we are adding to it a statement about how strongly we believe it. In this book, we develop a different way of producing hedged predictions. Aside from the elegance of our new methods, at least in comparison with the procedure that relies on a hold-out sample, the methods we develop have several important advantages.

As already mentioned in the preface, we do not have the rigid separation between learning and prediction, which is the feature of the traditional approaches that makes hedged prediction feasible. In our basic learning protocol learning and prediction are blended, yet our predictions are hedged.

Second, the hedged predictions produced by our new algorithms are much more confident and accurate. We have, of course, a different notion of a hedged prediction, so the comparison can be only informal; but the difference is so big that there is little doubt that the improvement is real from the practical point of view.

A third advantage of our methods is that the confidence with which the label of a new object is predicted is always tailored not only to the previously seen examples but also to that object.

1.3 The on-line transductive framework

The new methods presented in this book are quite general; they can be tried out, at least, in almost any problem of learning under randomness. The framework in which we introduce and study these methods is somewhat unusual, however. Most previous theoretical work in machine learning has been in an *inductive* and *off-line* framework: one uses a batch of old examples to produce a prediction rule, which is then applied to new examples. We begin instead with a framework that is *transductive*, in the sense advocated by Vapnik (1995, 1998), and *on-line*: one makes predictions sequentially, basing each new prediction on all the previous examples instead of repeatedly using a rule constructed from a fixed batch of examples.

On-line learning

Our framework is *on-line* because we assume that the examples are presented one by one. Each time, we observe the object and predict the label. Then we observe the label and go on to the next example. We start by observing the first object x_1 and predicting its label y_1 . Then we observe y_1 and the second object x_2 , and predict its label y_2 . And so on. At the n th step, we have observed the previous examples

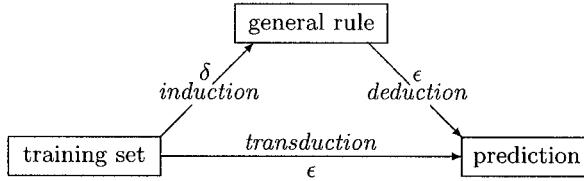


Fig. 1.3. Inductive and transductive prediction

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$$

and the new object x_n , and our task is to predict y_n . The quality of our predictions should improve as we accumulate more and more old examples. This is the sense in which we are learning.

Transduction

Vapnik's distinction between induction and transduction, as applied to the problem of prediction, is depicted in Fig. 1.3. In *inductive prediction* we first move from examples in hand to some more or less general rule, which we might call a prediction or decision rule, a model, or a theory; this is the *inductive step*. When presented with a new object, we derive a prediction from the general rule; this is the *deductive step*. In *transductive prediction*, we take a shortcut, moving from the old examples directly to the prediction about the new object.

Typical examples of the inductive step are estimating parameters in statistics and finding a “concept” (to use Valiant’s 1984 terminology) in statistical learning theory. Examples of transductive prediction are estimation of future observations in statistics (see, e.g., Cox and Hinkley 1974, §7.5) and nearest neighbors algorithms in machine learning.

In the case of simple predictions the distinction between induction and transduction is less than crisp. A method for doing transduction, in our online setting, is a method for predicting y_n from $x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n$. Such a method gives a prediction for any object that might be presented as x_n , and so it defines, at least implicitly, a rule, which might be extracted from $x_1, y_1, \dots, x_{n-1}, y_{n-1}$ (induction), stored, and then subsequently applied to x_n to predict y_n (deduction). So any real distinction is really at a practical and computational level: do we extract and store the general rule or not?

For hedged predictions the difference between transduction and induction goes deeper. We will typically want different notions of hedged prediction in the two frameworks. Mathematical results about induction typically involve two parameters, often denoted ϵ (the desired accuracy of the prediction rule) and δ (the probability of achieving the accuracy of ϵ), whereas results about transduction involve only one parameter, which we will denote ϵ (in this book,

the probability of error we are willing to tolerate); see Fig. 1.3. A detailed discussion can be found in Chap. 10, which also contains a historical perspective on the three main approaches to hedged prediction (inductive, Bayesian, and transductive).

On-line/off-line and transduction/induction compromises

When we work on-line, we would want to use a general rule extracted from $x_1, y_1, \dots, x_{n-1}, y_{n-1}$ only once, to predict y_n from x_n . After observing x_n and then y_n , we have a larger dataset, $x_1, y_1, \dots, x_n, y_n$, and we can use it to extract a new, possibly improved, general rule before trying to predict y_{n+1} from x_{n+1} . So from a purely conceptual point of view, induction seems silly in the on-line framework; it is more natural to say that we are doing transduction, even in cases where the general rule is easy to extract. As a practical matter, however, the computational cost of a transductive method may be high, and in this case, it may be sensible to compromise with the off-line or inductive approach. After accumulating a certain number of examples, we might extract a general rule and use it for a while, only updating it as frequently as is practical.

The methods we present in this book are most naturally described and are most amenable to mathematical analysis in the on-line framework. So we work out our basic theory in that framework, and this theory can be considered transductive. The theory extends, however, to the transductive/inductive compromise just described, where a general rule is extracted and used for a period of time before it is updated (see §4.1).

The theory also extends to relaxations of the on-line protocol that make it close to the off-line setting, and this is important, because most practical problems have at least some off-line aspects. If we are concerned with recognizing hand-written zip codes, for example, we cannot always rely on a human teacher to tell us the correct interpretation of each hand-written zip code; why not use such an *ideal teacher* directly for prediction? The relaxation of the on-line protocol considered in §4.3 includes “slow teachers”, who provide the feedback with a delay, and “lazy teachers”, who provide feedback only occasionally. In the example of zip codes recognition, this relaxation allows us to replace constant supervision by using returned letters for teaching or by occasional lessons.

1.4 Conformal prediction

Most of this book is devoted to a particular method that we call “conformal prediction”. When we use this method, we predict that a new object will have a label that makes it similar to the old examples in some specified way, and we use the degree to which the specified type of similarity holds within

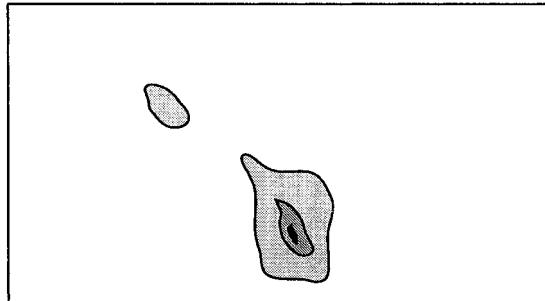


Fig. 1.4. An example of a nested family of prediction sets (casual prediction in black, confident prediction in dark gray, and highly confident prediction in light gray)

the old examples to estimate our confidence in the prediction. Our conformal predictors are, in other words, “confidence predictors”.

We need not explain here exactly how conformal prediction works. This is the topic of the next chapter. But we will explain informally what a confidence predictor aims to do and what it means for it to be valid and efficient.

Nested prediction sets

Suppose we want to pinpoint a target that lies somewhere within a rectangular field. This could be an on-line prediction problem; for each example, we predict the coordinates $y_n \in [a_1, a_2] \times [b_1, b_2]$ of the target from a set of measurements x_n .

We can hardly hope to predict the coordinates y_n exactly. But we can hope to have a method that gives a subset Γ_n of $[a_1, a_2] \times [b_1, b_2]$ where we can be confident y_n lies. Intuitively, the size of the *prediction set* Γ_n should depend on how great a probability of error we want to allow, and in order to get a clear picture, we should specify several such probabilities. We might, for example, specify the probabilities 1%, 5%, and 20%, corresponding to *confidence levels* 99%, 95%, and 80%. When the probability of the prediction set failing to include y_n is only 1%, we declare 99% confidence in the set (highly confident prediction). When it is 5%, we declare 95% confidence (confident prediction). When it is 20%, we declare 80% confidence (casual prediction). We might also want a 100% confidence set, but in practice this might be the whole field assumed at the outset to contain the target.

Figure 1.4 shows how such a family of prediction sets might look. The casual prediction pinpoints the target quite well, but we know that this kind of prediction can be wrong 20% of the time. The confident prediction is much bigger. If we want to be highly confident (make a mistake only for each 100th example, on average), we must accept an even lower accuracy; there is even a completely different location that we cannot rule out at this level of confidence.

In principle, a confidence predictor outputs prediction sets for all confidence levels, and these sets are nested, as in Fig. 1.4.

There are two important desiderata for a confidence predictor:

- They should be *valid*, in the sense that in the long run the frequency¹ of error does not exceed ϵ at each chosen confidence level $1 - \epsilon$.
- They should be *efficient*, in the sense that the prediction sets they output are as small as possible.

We would also like the predictor to be as conditional as possible – we want it to take full account of how difficult the particular example is.

Validity

Our **conformal predictors are always valid**. Fig. 1.5 shows the empirical confirmation of the validity for one particular conformal predictor that we study in Chap. 3. The solid, dash-dot and dotted lines show the cumulative number of errors for the confidence levels 99%, 95%, and 80%, respectively. As expected, the number of errors made grows linearly, and the slope is approximately 20% for the confidence level 80%, 5% for the confidence level 95%, and 1% for the confidence level 99%.

As we will see in Chap. 2, a precise discussion of the validity of conformal predictors actually requires that we distinguish **two kinds of validity: conservative and exact**. In general, a **conformal predictor is conservatively valid**: the probability it makes an error when it outputs a $1 - \epsilon$ set (i.e., a prediction set at a confidence level $1 - \epsilon$) is no greater than ϵ , and there is little dependence between errors it makes when predicting successive examples (at successive *trials*, as we will say). This implies, by the law of large numbers, that the long-run frequency of errors at confidence level $1 - \epsilon$ is about ϵ or less. In practice, the **conservativeness** is often not very great, especially when n is large, and so we get empirical results like those in Fig. 1.5, where the long-run frequency of errors is very close to ϵ . From a theoretical point of view, however, we must introduce a small element of deliberate randomization into the prediction process in order to get exact validity, where the probability of a $1 - \epsilon$ set being in error is exactly ϵ , errors are made independently at different trials, and the long-run frequency of errors converges to ϵ .

Efficiency

Machine learning has been mainly concerned with two types of problems:

- **Classification**, where the label space \mathbf{Y} is a small finite set (often binary).
- **Regression**, where the label space is the real line.

¹By “frequency” we usually mean “relative frequency”.

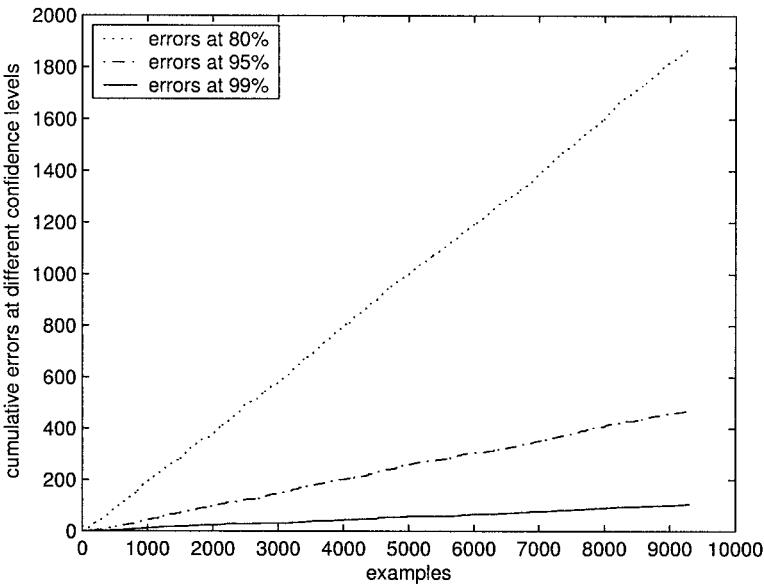


Fig. 1.5. On-line performance of a conformal predictor (“the 1-nearest neighbor conformal predictor”, described in Chap. 3) on the USPS data set (9298 handwritten digits, randomly permuted) for the confidence levels 80%, 95%, and 99%. The figures in this book are not too much affected by statistical variation (due to the random choice of the permutation of the data set)

In classification problems, a natural measure of efficiency of confidence predictors is the number of multiple predictions – the number of prediction sets containing two or more labels, at different confidence levels. In regression problems, the prediction set is often an interval of values, and a natural measure of efficiency of such a prediction is the length of the interval. In §2.3 we use the median length of the convex closures of prediction sets as a measure of efficiency of a sequence of predictions.

As we will see in Chap. 2, there are many conformal predictors for any particular on-line prediction problem, whether it is a classification problem or a regression problem. Indeed, we can construct a conformal predictor from any method for scoring the similarity (conformity, as we call it) of a new example to old ones. All these conformal predictors, it turns out, are valid. Which is most efficient – which produces the smallest prediction sets in practice – will depend on details of the environment that we may not know in advance.

In Chap. 3 we show that there exists a “universal” randomized conformal predictor, making asymptotically no more multiple predictions than any valid confidence predictor. This asymptotic result may, however, have limited relevance to practical prediction problems (as discussed in the next section).

Conditionality

The goal of conditionality can be explained with a simple example discussed by David Cox (1958b). Suppose there are two categories of objects, “easy” (easy to predict) and “hard” (hard to predict). We can tell which objects belong to which category, and the two categories occur with equal probability; about 50% of the objects we encounter are easy, and 50% hard. We have a prediction method that applies to all objects, hard and easy, and has error rate 5%. We do not know what the error rate is for hard objects, but perhaps it is 8%, and we get an overall error rate of 5% only because the rate for easy objects is 2%. In this situation, we may feel uncomfortable, when we encounter a hard object, about appealing to the average error rate of 5% and saying that we are 95% confident of our prediction.

Whenever there are features of objects that we know make the prediction easier or harder, we would like to take these features into account – to condition on them. This is done by conformal predictors almost automatically: they are designed for specific applications so that their predictions take fullest possible account of the individual object to be predicted. What is not achieved automatically is the validity separately for hard and easy objects. It is possible, for example, that if a figure such as Fig. 1.5 were constructed for easy objects only, or for hard objects only, the slopes of the cumulative error lines would be different. We would get the correct slope if we average the slope for easy objects and the slope for hard objects, but we would ideally like to have the “conditional validity”: validity for both categories of objects. As we show in §4.5, this can be achieved by modifying the definition of conformal predictors. In fact, the conditional validity is handled by a general theory that also applies when we segregate examples not by their difficulty but by their time of arrival, as when we are using an inductive rule that we update only at specified intervals.

Flexibility of conformal predictors

A useful feature of our method is that a conformal predictor can be built on top of almost any machine-learning algorithm. The latter, which we call the *underlying algorithm*, may produce hedged predictions, simple predictions, or simple predictions complemented by ad hoc measures of confidence; our experience is that it is always possible to transform it into a conformal predictor that inherits its predictive performance but is, of course, valid, just like any other conformal predictor. In this book we explain how to build conformal predictors using such methods as nearest neighbors, support vector machines, bootstrap, boosting, neural networks, decision trees, ridge regression, logistic regression, and any Bayesian algorithm (see §§2.3, 3.1, 4.2).

1.5 Probabilistic prediction under unconstrained randomness

There are many ways to do classification and regression under unconstrained randomness and for high-dimensional examples. Conformal predictors, for example, combine good theoretical properties with high accuracy in practical problems. It is true that the environment has to be benign, in some sense, for any learning method to be successful, but there are no obvious insurmountable barriers for classification and regression. The situation changes if we move to the harder problem of *probabilistic prediction*: that of guessing the probability distribution for the new object's label. Features of data that can reasonably be expected in typical machine-learning applications become such barriers.

For simplicity, we will assume in this section that the label is binary, 0 or 1. In this case the probabilistic prediction for the label of the new object boils down to one number, the predicted probability that the label is 1.

The problem of probabilistic prediction is discussed in Chaps. 5, 6, and 9. Probabilistic prediction is impossible in an important sense, but there are also senses in which it is possible. So this book gives more than one answer to the question “Is probabilistic prediction possible?” We start with a “yes” answer.

Universally consistent probabilistic predictor

Stone (1977) showed that a nearest neighbors probabilistic predictor (whose probabilistic prediction is the fraction of objects classified as 1 among the k nearest neighbors of the new object, with a suitably chosen k) is *universally consistent*, in the sense that the difference between the probabilistic prediction and the true conditional probability given the object that the label is 1 converges to zero in probability. The only essential assumption is randomness²; there are no restrictive regularity conditions.

Stone's actual result was more general, and it has been further extended in different directions. One of these extension is used in Chap. 3 for constructing a universal randomized conformal predictor.

Probabilistic prediction using a finite data set

The main obstacle in applying Stone's theorem is that the convergence it asserts is not uniform. The situation that we typically encounter in practice is that we are given a set of examples and a new object and we would like to estimate the probability that the label of the new object is 1. It is well understood that in this situation the applicability of Stone's theorem is very

²The other assumption made by Stone was that the objects were coming from a Euclidean space; since “Euclidean” is equivalent to “Borel” in the context of existence of a universally consistent probabilistic predictor, this assumption is very weak.

limited (see, e.g., Devroye et al. 1996, §7.1). In Chap. 5 we give a new, more direct, formalization of this observation.

We say that a data set consisting of old examples and one new object is *diverse* if no object in it is repeated (in particular, the new object is different from all old objects). The main result of §5.2 asserts that any nontrivial (not empty and not containing 0 and 1) prediction interval for the conditional probability given the new object that the new label is 1 is inadmissible if the data set is diverse and randomness is the only assumption.

The assumption that the data set is diverse is related to the assumption of a high-dimensional environment. If the objects are, for example, complex images, we will not expect precise repetitions among them.

Venn prediction

The results of Chap. 5 show that it is impossible to estimate the true conditional probabilities under the conditions stated; that chapter also contains a result that it is impossible to find conditional probabilities that are as good (in the sense of the algorithmic theory of randomness) as the true probabilities. If, however, we are prepared to settle for less and only want probabilities that are “well calibrated” (in other words, have a frequentist justification), a modification of conformal predictors which we call Venn predictors will achieve this goal, in a very strong non-asymptotic sense. This is the subject of Chap. 6, which is one of the longest in this book. The main problem that we have to deal with in this chapter is that one cannot guarantee that miscalibration will not happen: everything can happen (perhaps with a small probability) for finite sequences and typical probability distributions. But in the case of Venn predictors, any evidence against calibration translates into evidence, at least as strong, against the assumption of randomness; therefore, we expect Venn predictors to be well calibrated as long as we accept the hypothesis of randomness. A significant part of the chapter is devoted to the ways of testing calibration and randomness.

1.6 Beyond randomness

In this book we also consider testing the assumption of randomness and alternatives to this assumption. The most radical alternative is introduced in Chaps. 8 and 9 under the name of “on-line compression modeling”.

Testing randomness

This is the topic of Chap. 7. We start it by adapting the mathematical apparatus developed in the previous chapters to testing the assumption of randomness. The usual statistical approach to testing (sometimes called

the “Neyman–Pearson–Wald” theory) is essentially off-line: in the original Neyman–Pearson approach (see, e.g., Lehmann 1986), the sample size is chosen *a priori*, and in Wald’s (1947) sequential analysis, the sample size is data-dependent but still at some point a categorical decision on whether the null hypothesis is rejected or not is taken (with probability one). The approach of §7.1 is on-line: we constantly update the strength of evidence against the null hypothesis of randomness. Finding evidence against the null hypothesis involves gambling against it, and the strength of evidence equals the gambler’s current capital. For further details and the history of this approach to testing, see Shafer and Vovk 2001. The main mathematical finding of §7.1 is that there exists a wide family of “exchangeability martingales”, which can be successfully applied to detecting lack of randomness.

Low-dimensional dynamic models

The ability to test the assumption of randomness immediately provides opportunities for extending the range of stochastic environments to which one can apply the idea of conformal prediction. In §7.2 we consider the simple case where we are given a parametric family of transformations one of which is believed to transform the observed data sequence into a random sequence. If the parameter is a vector in a low-dimensional linear space, we can often hope to be able to detect lack of randomness of the transformed data sequence for most values of the parameter as the number of observed examples grows. When the range of possible values of the parameter becomes very narrow, conformal prediction can be used.

Islands of randomness

When we are willing to make the assumption of randomness, or some version of this assumption as described in the previous subsection, about a data sequence, it usually means that this data sequence was obtained from a much bigger sequence by careful filtering. When observing the real world around us, we cannot hope that a simple model such as randomness will explain much, but the situation changes if we, e.g., discard all observations except the results of fair coin tosses.

In §7.3, we briefly discuss the case where randomness can appear as a property of only relatively small subsequences of the full data sequence. Such a “big picture” is of great interest to philosophers (see, e.g., Venn 1866). Once we know that some subsequence is random (this knowledge can be based on an initial guess and then using as severe tests as we can think of to try and falsify this guess; §7.1 provides the means for the second stage), we can apply the theory developed under the assumption of randomness to make predictions.

On-line compression models

As we will see in Chap. 8, the idea of conformal prediction generalizes from learning under randomness, where examples are independent and identically distributed, to “on-line compression models”. In an on-line compression model, it is assumed that the data can be summarized in way that can be updated as new examples come in, and the only probabilities given are backward probabilities – probabilities for how the updated summary might have been obtained.

On-line compression models derive from the work of Andrei Kolmogorov. They open a new direction for broadening the applicability of machine-learning methods, giving a new meaning to the familiar idea that learning can be understood as information compression.

In Chap. 8 we consider in detail three important on-line compression models (Gaussian, Markov, exchangeability) and their variants. In Chap. 9 we extend the idea of Venn prediction to on-line compression modeling and apply it to a new model, which we call the “hypergraphical model”.

1.7 Bibliographical remarks

Each chapter of this book ends with a section entitled “Bibliographical remarks”, or similarly. These sections are set in a small font and may use mathematical notions and results not introduced elsewhere in the book.

Turing suggested the idea of machine learning in his paper published in *Mind* as an approach to solving his famous “imitation game” (Turing 1950, §1).

A recent empirical study of various bounds on prediction accuracy is reported in Langford 2004. It found the hold-out estimate to be a top performer.

Mitchell (1997, §8.6) discusses advantages and disadvantages of inductive and transductive approaches to making simple predictions. The near-synonyms for “transductive learning” used in that book are “lazy learning” and “instance-based learning”.

2

Conformal prediction

In this chapter we formally introduce conformal predictors. After giving the necessary definitions, we will prove that when a conformal predictor is used in the on-line mode, its output is valid, not only in the asymptotic sense that the sets it predicts for any fixed confidence level $1 - \epsilon$ will be wrong with frequency at most ϵ (approaching ϵ in the case of smoothed conformal predictors) in the long run, but also in a much more precise sense: the error probability of a smoothed conformal predictor is ϵ at every trial and errors happen independently at different trials. In §2.4 we will see that conformal prediction is indispensable for achieving this kind of validity. The basic procedure of conformal prediction might look computationally inefficient when the label set is large, but in §2.3 we show that in the case of, e.g., least squares regression (where the label space \mathbb{R} is uncountable) there are ways of making conformal predictors much more efficient.

2.1 Confidence predictors

The conformal predictors we define in this chapter are confidence predictors – they make a range of successively more specific predictions with successively less confidence. In this section we define precisely what we mean by a confidence predictor and its validity.

Assumptions

We assume that Reality outputs successive pairs

$$(x_1, y_1), (x_2, y_2), \dots , \quad (2.1)$$

called *examples*. Each example (x_i, y_i) consists of an *object* x_i and its *label* y_i . The objects are elements of a measurable space \mathbf{X} called the *object space* and the labels are elements of a measurable space \mathbf{Y} called the *label space*.

We assume that \mathbf{X} is non-empty and that \mathbf{Y} contains at least two essentially different elements¹. When we need a more compact notation, we write z_i for (x_i, y_i) . We set

$$\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$$

and call \mathbf{Z} the *example space*. Thus the infinite data sequence (2.1) is an element of the measurable space \mathbf{Z}^∞ .

When we say that the objects are *absent*, we mean that $|\mathbf{X}| = 1$. In this case x_i do not carry any information and do not need to be mentioned; we will then identify \mathbf{Y} and \mathbf{Z} .

Our standard assumption is that Reality chooses the examples independently from some probability distribution Q on \mathbf{Z} – i.e., that the infinite sequence z_1, z_2, \dots is drawn from the *power probability distribution* Q^∞ in \mathbf{Z}^∞ . Most of the results of this book hold under this *randomness assumption*, but usually we need only the slightly weaker assumption that the infinite data sequence (2.1) is drawn from a distribution P on \mathbf{Z}^∞ that is *exchangeable*. The statement that P is exchangeable means that for every positive integer n , every permutation π of $\{1, \dots, n\}$, and every measurable set $E \subseteq \mathbf{Z}^n$,

$$\begin{aligned} P \{(z_1, z_2, \dots) \in \mathbf{Z}^\infty : (z_1, \dots, z_n) \in E\} \\ = P \{(z_1, z_2, \dots) \in \mathbf{Z}^\infty : (z_{\pi(1)}, \dots, z_{\pi(n)}) \in E\} . \end{aligned}$$

Every power distribution is exchangeable, and under a natural regularity condition (\mathbf{Z} is a Borel space), any exchangeable distribution on \mathbf{Z}^∞ is a mixture of power distributions; for details, see §A.5. In our mathematical results, we usually use the randomness assumption or the exchangeability assumption depending on which one leads to a stronger statement.

Simple predictors and confidence predictors

We assume that at the n th trial Reality first announces the object x_n and only later announces the label y_n . If we simply want to predict y_n , then we need a function

$$D : \mathbf{Z}^* \times \mathbf{X} \rightarrow \mathbf{Y} . \tag{2.2}$$

We call such a function a *simple predictor*, always assuming it is measurable. For any sequence of old examples, say $x_1, y_1, \dots, x_{n-1}, y_{n-1} \in \mathbf{Z}^*$, and any new object, say $x_n \in \mathbf{X}$, it gives $D(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \in \mathbf{Y}$ as its prediction for the new label y_n .

As we explained in §1.4, however, we have a more complicated notion of prediction. Instead of merely choosing a single element of \mathbf{Y} as our prediction

¹Formally, the σ -algebra on \mathbf{Y} is assumed to be different from $\{\emptyset, \mathbf{Y}\}$. It is convenient to assume that for each pair of distinct elements of \mathbf{Y} there is a measurable set containing only one of them; we will do this without loss of generality, and then our assumption about \mathbf{Y} is that $|\mathbf{Y}| > 1$.

for y_n , we want to give a range of more or less precise predictions, each labeled with a degree of confidence. We want to give subsets of \mathbf{Y} large enough that we can be confident that y_n will fall in them, while also giving smaller subsets in which we are less confident. An algorithm that predicts in this sense requires an additional input $\epsilon \in (0, 1)$, which we call the *significance level*; the complementary value $1 - \epsilon$ is called the *confidence level*. Given all these inputs, say

$$x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, \epsilon,$$

an algorithm of the type that interests us, say Γ , outputs a subset

$$\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$$

of \mathbf{Y} . (We position ϵ as a superscript instead of placing it with the other arguments.) We require this subset to shrink as ϵ is increased: Γ must satisfy

$$\Gamma^{\epsilon_1}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma^{\epsilon_2}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \quad (2.3)$$

whenever $\epsilon_1 \geq \epsilon_2$. Intuitively, once we observe the *incomplete data sequence*

$$x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n \quad (2.4)$$

and chose the significance level ϵ , Γ predicts that

$$y_n \in \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n), \quad (2.5)$$

and the smaller ϵ is the more confident the prediction is. According to condition (2.3), we are more confident in less specific predictions.

Formally, we call a measurable function

$$\Gamma : \mathbf{Z}^* \times \mathbf{X} \times (0, 1) \rightarrow 2^{\mathbf{Y}} \quad (2.6)$$

($2^{\mathbf{Y}}$ is the set of all subsets of \mathbf{Y}) that satisfies (2.3), for all significance levels $\epsilon_1 \geq \epsilon_2$, all positive integers n , and all incomplete data sequences (2.4), a *confidence predictor* (or *deterministic confidence predictor*). The requirement that Γ be measurable means that for each n the set of sequences $\epsilon, x_1, y_1, \dots, x_n, y_n$ satisfying (2.5) is a measurable subset of $(0, 1) \times (\mathbf{X} \times \mathbf{Y})^n$.

Validity

Let us begin with an intuitive explanation of the notions of exact and conservative validity. For each significance level ϵ , we want to have confidence $1 - \epsilon$ in our prediction (2.5) about y_n . This means that the probability of the prediction being in error – the probability of the event (2.5) not happening – should be ϵ . Moreover, since we are making a whole sequence of predictions, first for y_1 , then for y_2 , and so on, we would like these error events to be independent. If these conditions are met no matter what *exchangeable probability*

distribution P governs the sequence of examples, then we say that the confidence predictor Γ is “exactly valid”, or, more briefly, “exact”. This means that making errors with Γ is like getting heads when making independent tosses of a biased coin whose probability of heads is always ϵ . If the probabilities for errors are allowed to be even less than this, then we say that the confidence predictor Γ is “conservatively valid” or, more briefly, “conservative”.

When we make independent tosses of a biased coin whose probability of heads is always ϵ , the frequency of heads will converge to ϵ with probability one – this is the law of large numbers. So the frequency of errors at significance level ϵ for an exactly valid confidence predictor converges to ϵ with probability one. As we will see, confidence predictors sometimes have this asymptotic property even when they are not exactly valid. So we give the property a name of its own; we call a confidence predictor that has it “asymptotically exact”. Similarly, we call a confidence predictor for which the frequency of errors is asymptotically no more than ϵ (for which the upper limit of the frequency of errors is at most ϵ) with probability one “asymptotically conservative”.

In order to restate these definitions in a way sufficiently precise to exclude possible misunderstandings, we now introduce a formal notation for the errors Γ makes when it processes the data sequence

$$\omega = (x_1, y_1, x_2, y_2, \dots) \quad (2.7)$$

at significance level ϵ . Whether Γ makes an error on the n th trial can be represented by a number that is one in case of an error and zero in case of no error:

$$\text{err}_n^\epsilon(\Gamma, \omega) := \begin{cases} 1 & \text{if } y_n \notin \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \\ 0 & \text{otherwise,} \end{cases} \quad (2.8)$$

and the number of errors during the first n trials is

$$\text{Err}_n^\epsilon(\Gamma, \omega) := \sum_{i=1}^n \text{err}_i^\epsilon(\Gamma, \omega). \quad (2.9)$$

It is enlightening to think about the error counts err_n^ϵ and Err_n^ϵ in the context of the protocol followed as the examples are presented and the predictions are made:

```

 $\text{Err}_0^\epsilon := 0$  for all  $\epsilon \in (0, 1)$ ;  

FOR  $n = 1, 2, \dots$  :  

    Reality outputs  $x_n \in \mathbf{X}$ ;  

    Predictor outputs  $\Gamma_n^\epsilon \subseteq \mathbf{Y}$  for all  $\epsilon \in (0, 1)$ ;  

    Reality outputs  $y_n \in \mathbf{Y}$ ;  

     $\text{err}_n^\epsilon := \begin{cases} 1 & \text{if } y_n \notin \Gamma_n^\epsilon \\ 0 & \text{otherwise} \end{cases}$  for all  $\epsilon \in (0, 1)$ ;  

     $\text{Err}_n^\epsilon := \text{Err}_{n-1}^\epsilon + \text{err}_n^\epsilon$  for all  $\epsilon \in (0, 1)$   

END FOR.

```

This is a game protocol, and we can consider strategies for both players, Reality and Predictor. We have assumed that Reality's strategy is randomized; her moves are generated from an exchangeable probability distribution P on \mathbf{Z}^∞ . A confidence predictor Γ is, by definition, a measurable strategy for Predictor.

If ω is drawn from an exchangeable probability distribution P , the number $\text{err}_n^\epsilon(\Gamma, \omega)$ is the realized value of a random variable, which we may designate $\text{err}_n^\epsilon(\Gamma, P)$. Formally, the confidence predictor Γ is *exactly valid* if for each ϵ ,

$$\text{err}_1^\epsilon(\Gamma, P), \text{err}_2^\epsilon(\Gamma, P), \dots \quad (2.10)$$

is a sequence of independent Bernoulli random variables with parameter ϵ – i.e., if it is a sequence of independent random variables each of which has probability ϵ of being one and probability $1 - \epsilon$ of being zero – no matter what exchangeable distribution P we draw ω from. Unfortunately, the notion of exact validity is vacuous for confidence predictors.

Theorem 2.1. *No confidence predictor is exactly valid.*

The notion of conservative validity is more complex; now we only require that $\text{err}_n^\epsilon(\Gamma, P)$ be *dominated in distribution* by a sequence of independent Bernoulli random variables with parameter ϵ . Formally, the confidence predictor Γ is *conservatively valid* if for any exchangeable probability distribution P on \mathbf{Z}^∞ there exists a probability space with two families

$$(\xi_n^{(\epsilon)} : \epsilon \in (0, 1), n = 1, 2, \dots), \quad (\eta_n^{(\epsilon)} : \epsilon \in (0, 1), n = 1, 2, \dots)$$

of $\{0, 1\}$ -valued random variables such that:

- for a fixed ϵ , $\xi_1^{(\epsilon)}, \xi_2^{(\epsilon)}, \dots$ is a sequence of independent Bernoulli random variables with parameter ϵ ;
- for all n and ϵ , $\eta_n^{(\epsilon)} \leq \xi_n^{(\epsilon)}$;
- the joint distribution of $\text{err}_n^\epsilon(\Gamma, P)$, $\epsilon \in (0, 1)$, $n = 1, 2, \dots$, coincides with the joint distribution of $\eta_n^{(\epsilon)}$, $\epsilon \in (0, 1)$, $n = 1, 2, \dots$.

(It might have been natural to also require that $\xi_n^{(\epsilon_1)} \geq \xi_n^{(\epsilon_2)}$ whenever $\epsilon_1 \geq \epsilon_2$, but it is easy to check that the inclusion of this condition leads to an equivalent definition.)

To conclude, we define *asymptotic validity*. The confidence predictor Γ is *asymptotically exact* if for any exchangeable probability distribution P on \mathbf{Z}^∞ generating examples z_1, z_2, \dots and any significance level ϵ ,

$$\lim_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon(\Gamma, P)}{n} = \epsilon \quad (2.11)$$

with probability one. It is *asymptotically conservative* if for any exchangeable probability distribution P on \mathbf{Z}^∞ and any significance level ϵ ,

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon(\Gamma, P)}{n} \leq \epsilon \quad (2.12)$$

with probability one.

Proposition 2.2. *An exact confidence predictor is asymptotically exact. A conservative confidence predictor is asymptotically conservative.*

This proposition is an immediate consequence of the law of large numbers.

Randomized confidence predictors

We will also be interested in randomized confidence predictors, which depend, additionally, on elements of an auxiliary probability space. The main advantage of randomization in this context is that, as we will see, there are many randomized confidence predictors that are exactly valid. Formally, we define a *randomized confidence predictor* to be a measurable function

$$\Gamma : (\mathbf{X} \times [0, 1] \times \mathbf{Y})^* \times (\mathbf{X} \times [0, 1]) \times (0, 1) \rightarrow 2^\mathbf{Y} \quad (2.13)$$

which, for all significance levels $\epsilon_1 \geq \epsilon_2$, all positive integer n , and all incomplete data sequences

$$x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n,$$

where $x_i \in \mathbf{X}$, $\tau_i \in [0, 1]$ and $y_i \in \mathbf{Y}$ for all i , satisfies

$$\begin{aligned} \Gamma^{\epsilon_1}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \\ \subseteq \Gamma^{\epsilon_2}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n). \end{aligned}$$

We will always assume that τ_1, τ_2, \dots are random variables that are independent (between themselves and of anything else) and distributed uniformly in $[0, 1]$; this is what one expects to get from a random number generator.

We define $\text{err}_n^\epsilon(\Gamma, \omega)$ and $\text{Err}_n^\epsilon(\Gamma, \omega)$ as before, only now they also depend on the τ 's. In other words, $\text{err}_n^\epsilon(\Gamma, \omega)$ is defined by (2.8) with x_i now being *extended objects* $x_i \in \mathbf{X} \times [0, 1]$; $\text{Err}_n^\epsilon(\Gamma, \omega)$ is defined, as before, by (2.9).

In general, many definitions for randomized confidence predictors are special cases of the corresponding definitions for deterministic confidence predictors: the latter should just be applied to the *extended object space* $\mathbf{X} \times [0, 1]$ (whose elements consist of both the object and the random number to be used at a given trial).

If Γ is a randomized confidence predictor, P is an exchangeable distribution on \mathbf{Z}^∞ , and $n \in \mathbb{N}$,

$$\text{err}_n^\epsilon(\Gamma, (x_1, \tau_1, y_1, x_2, \tau_2, y_2, \dots))$$

and

$$\text{Err}_n^\epsilon(\Gamma, (x_1, \tau_1, y_1, x_2, \tau_2, y_2, \dots))$$

where $(x_1, y_1), (x_2, y_2), \dots$ are drawn from P and τ_1, τ_2, \dots are drawn independently from \mathbf{U}^∞ , the uniform distribution on $[0, 1]^\infty$, are random variables, which will be denoted $\text{err}_n^\epsilon(\Gamma, P)$ and $\text{Err}_n^\epsilon(\Gamma, P)$, respectively. We say that Γ is *exactly valid* if for each $\epsilon \in (0, 1)$, (2.10) is a sequence of independent Bernoulli random variables with parameter ϵ .

We are not really interested in the notion of conservative validity for randomized confidence predictors.

2.2 Conformal predictors

We start by defining the concept of a **nonconformity measure**. Intuitively, this is a way of measuring how different a new example is from old examples. There are many different nonconformity measures, and each one defines a conformal predictor and a smoothed conformal predictor.

Bags

The order in which old examples appear should not make any difference, and in order to formalize this point, we need the concept of a bag (also called a multiset). A *bag* of size $n \in \mathbb{N}$ is a collection of n elements some of which may be identical; a bag resembles a set in that the order of its elements is irrelevant, but it differs from a set in that repetition is allowed. To identify a bag, we must say what elements it contains and how many times each of these elements is repeated. We write $\{z_1, \dots, z_n\}$ for the bag consisting of elements z_1, \dots, z_n , some of which may be identical with each other.

Although the elements of a bag are not ordered, there is an ordering in our notation, and we will make use of it. The bag $\{z_1, \dots, z_{20}\}$ is the bag we get from z_1, \dots, z_{20} when we ignore their order, but because we have identified the bag using the elements in a certain order, we can manipulate it using our knowledge of this order. We can, for example, talk about the bag we get when we remove z_6 (while leaving any other z_i that might be equal to z_6); this is a bag of size 19 – the bag $\{z_1, \dots, z_5, z_7, \dots, z_{20}\}$.

We write $\mathbf{Z}^{(n)}$ for the set of all bags of size n of elements of a measurable space \mathbf{Z} . The set $\mathbf{Z}^{(n)}$ is itself a measurable space. It can be defined formally as the power space \mathbf{Z}^n with a nonstandard σ -algebra, consisting of measurable subsets of \mathbf{Z}^n that contain all permutations of their elements. We write $\mathbf{Z}^{(*)}$ for the set of all bags of elements of \mathbf{Z} (the union of all the $\mathbf{Z}^{(n)}$).

Nonconformity and conformity

As we have already said, a nonconformity measure is a way of scoring how different a new example is from a bag of old examples. Formally, a *nonconformity measure* is a measurable mapping

$$A : \mathbf{Z}^{(*)} \times \mathbf{Z} \rightarrow \overline{\mathbb{R}} ;$$

to each possible bag of old examples and each possible new example, A assigns a numerical score indicating how different the new example is from the old ones.

It is easy to invent nonconformity measures, especially when we already have methods for prediction at hand:

- If the examples are merely numbers ($\mathbf{Z} = \mathbb{R}$) and it is natural to take the average of the old examples as the simple predictor of the new example, then we might define the **nonconformity of a new example** as the **absolute value of its difference from the average of the old examples**. Alternatively, we could use instead the absolute value of its difference from the median of the old examples.
- In a regression problem where examples are pairs of numbers, say $z_i = (x_i, y_i)$, we might define the **nonconformity of a new example (x, y)** as **the absolute value of the difference between y and the predicted value \hat{y}** calculated from x and the old examples.

We will discuss this way of detecting nonconformity further at the end of this section. But whether a particular function on $\mathbf{Z}^{(*)} \times \mathbf{Z}$ is an appropriate way of measuring nonconformity will always be open to discussion, and we do not need to enter into this discussion at this point. In our general theory, we will call *any* measurable function on $\mathbf{Z}^{(*)} \times \mathbf{Z}$ taking values in the extended real line a nonconformity measure.

Given a nonconformity measure A , a sequence z_1, \dots, z_l of examples, and an example z , we can score how different z is from the bag $\{z_1, \dots, z_l\}$; namely, $A(\{z_1, \dots, z_l\}, z)$ is called the *nonconformity score* for z .

Of course, instead of looking at functions that we feel measure nonconformity, we could look at functions that we feel measure conformity. We call such a function, say B , a *conformity measure*, and we can use it to define *conformity scores* $B(\{z_1, \dots, z_l\}, z)$. Formally, a *conformity measure* is a measurable function of the type $\mathbf{Z}^{(*)} \times \mathbf{Z} \rightarrow \overline{\mathbb{R}}$ (so there is no difference between conformity measures and nonconformity measures as mathematical objects). If we begin in this way, then nonconformity appears as a derivative idea. Given a *conformity measure* B we can define a nonconformity measure A using any strictly decreasing transformation, say $A := -B$ or perhaps (if B takes only positive values) $A := 1/B$. Given our goal, prediction, beginning with conformity might seem the more natural approach. As we will explain shortly, our strategy for prediction is to predict that a new label will be among the labels that best make a new example conform with old examples, and it is more natural to emphasize the labels that we include in the prediction (the most conforming ones) rather than the labels that we exclude (the most nonconforming ones). But in practice, it is often more natural to begin with nonconformity measures. For example, when we compare a new example with an average of old examples, we will usually first define a distance between the two rather

than devise a way to measure their closeness. For this reason, we emphasize nonconformity rather than conformity. Doing so is consistent with tradition in mathematical statistics, where test statistics are usually defined so as to measure discrepancy rather than agreement.

We sometimes find it convenient to consider separately how a nonconformity measure deals with bags of different sizes: if A is a nonconformity measure, for each $n = 1, 2, \dots$ we define the function

$$A_n : \mathbf{Z}^{(n-1)} \times \mathbf{Z} \rightarrow \bar{\mathbb{R}} \quad (2.14)$$

as the restriction of A to $\mathbf{Z}^{(n-1)} \times \mathbf{Z}$. The sequence $(A_n : n \in \mathbb{N})$, which we abbreviate to (A_n) when there is no danger of confusion, will also be called a nonconformity measure. Analogous conventions will be used for conformity measures.

p-values

Given a nonconformity measure (A_n) and a bag $\{z_1, \dots, z_n\}$, we can compute the nonconformity score

$$\alpha_i := A_n(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}, z_i) \quad (2.15)$$

for each example z_i in the bag. Because a nonconformity measure (A_n) may be scaled however we like, the numerical value of α_i does not, by itself, tell us how unusual (A_n) finds z_i to be. For that, we need a comparison of α_i to the other α_j . A convenient way of making this comparison is to compute the fraction

$$\frac{|\{j = 1, \dots, n : \alpha_j \geq \alpha_i\}|}{n}. \quad (2.16)$$

This fraction, which lies between $1/n$ and 1, is called the *p-value* for z_i . It is the fraction of the examples in the bag as nonconforming as z_i . If it is small (close to its lower bound $1/n$ for a large n), then z_i is very nonconforming (an outlier). If it is large (close to its upper bound 1), then z_i is very conforming.

If we begin with a conformity measure (B_n) rather than a nonconformity measure, then we can define the p-value for z_i by

$$\frac{|\{j = 1, \dots, n : \beta_j \leq \beta_i\}|}{n},$$

where the β_j are the conformity scores. This gives the same result as we would obtain from (2.16) using a nonconformity measure (A_n) obtained from (B_n) by means of a strictly decreasing transformation.

Definition of conformal predictors

Every nonconformity measure determines a confidence predictor. Given a new object x_n and a level of significance, this predictor provides a prediction set

that should contain the object's label y_n . We obtain the set by supposing that y_n will have a value that makes (x_n, y_n) conform with the previous examples. The level of significance determines the amount of conformity (as measured by the p-value) that we require.

Formally, the *conformal predictor determined by a nonconformity measure* (A_n) is the confidence predictor Γ obtained by setting

$$\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \quad (2.17)$$

equal to the set of all labels $y \in \mathbf{Y}$ such that

$$\frac{|\{i = 1, \dots, n : \alpha_i \geq \alpha_n\}|}{n} > \epsilon, \quad (2.18)$$

where

$$\begin{aligned} \alpha_i &:= A_n(\ell(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_{n-1}, y_{n-1}), \\ &\quad (x_n, y) \rfloor, (x_i, y_i)), \quad i = 1, \dots, n-1, \\ \alpha_n &:= A_n(\ell(x_1, y_1), \dots, (x_{n-1}, y_{n-1}) \rfloor, (x_n, y)). \end{aligned} \quad (2.19)$$

In general, a *conformal predictor* is a conformal predictor determined by some nonconformity measure.

The left-hand side of (2.18) is the p-value of (x_n, y) in the bag consisting of it and the old examples (cf. (2.16)). So our prediction with significance level ϵ (or confidence level $1 - \epsilon$) is that the value of y_n will make (x_n, y_n) have a p-value greater than ϵ when it is bagged with the old examples. We are 98% confident, for example, that we will get a value for y_n that gives (x_n, y_n) a p-value greater than 0.02. In other words, we are 98% confident that

$$\frac{\text{number of examples that conform worse or the same as } (x_n, y_n)}{n}$$

will exceed 0.02.

If A is a conformity measure, the conformal predictor determined by A is defined by (2.18) with “ \geq ” replaced by “ \leq ”.

Validity

Proposition 2.3. *All conformal predictors are conservative.*

It follows by Proposition 2.2 that a conformal predictor is asymptotically conservative. Of course, more can be said. Using the law of the iterated logarithm instead of the law of large numbers, we can strengthen (2.12) to

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon(\Gamma, \omega) - n\epsilon}{\sqrt{2\epsilon(1-\epsilon)n \ln \ln n}} \leq 1.$$

We will also state two finite-sample implications of Proposition 2.3: Hoeffding's inequality (see p. 287) implies that, for any positive integer N and any constant $\delta > 0$,

$$P \{ \omega : \text{Err}_N^\epsilon(\Gamma, \omega) \geq N(\epsilon + \delta) \} \leq e^{-2N\delta^2};$$

the central limit theorem implies that, for any constant c ,

$$\limsup_{N \rightarrow \infty} P \left\{ \omega : \text{Err}_N^\epsilon(\Gamma, \omega) \geq N\epsilon + c\sqrt{N} \right\} \leq \frac{1}{\sqrt{2\pi}} \int_{\frac{c}{\sqrt{\epsilon(1-\epsilon)}}}^{\infty} e^{-u^2/2} du.$$

For a graphical illustration of asymptotic conservativeness, see Fig. 1.5 on p. 10.

Smoothed conformal predictors

In this section we introduce a modification of conformal predictors which will allow us to simplify and strengthen Proposition 2.3. The *smoothed conformal predictor determined by the nonconformity measure* (A_n) is the following randomized confidence predictor Γ : the set

$$\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n)$$

consists of the $y \in \mathbf{Y}$ satisfying

$$\frac{|\{i = 1, \dots, n : \alpha_i > \alpha_n\}| + \tau_n |\{i = 1, \dots, n : \alpha_i = \alpha_n\}|}{n} > \epsilon, \quad (2.20)$$

where α_i are defined, as before, by (2.19). The left-hand side of (2.20) is called the *smoothed p-value*.

The main difference of (2.20) from (2.18) is that in the former we treat the borderline cases $\alpha_i = \alpha_n$ more carefully. Instead of increasing the p-value by $1/n$ for each $\alpha_i = \alpha_n$, we increase it by a random amount between 0 and $1/n$.

When n is not too small, it is typical for almost all $\alpha_1, \dots, \alpha_n$ to be different, and then there is very little difference between conformal predictors and smoothed conformal predictors.

Proposition 2.4. *Any smoothed conformal predictor is exactly valid.*

This proposition will be proved in Chap. 8 (as a special case of Theorem 8.1 on p. 193). It immediately implies Proposition 2.3: if a smoothed conformal predictor Γ and a conformal predictor Γ^\dagger are constructed from the same nonconformity measure, the latter's errors err_n^\dagger never exceed the former's errors err_n , $\text{err}_n^\dagger \leq \text{err}_n$. It also implies

Corollary 2.5. *Every smoothed conformal predictor is asymptotically exact.*

A general scheme for defining nonconformity

There are many different ways of defining nonconformity measures; we will consider them more systematically in the following two chapters, and here we will only explain the most basic approach, which in the next section will be illustrated in the case of regression, $\mathbf{Y} = \mathbb{R}$.

Suppose we are given a bag $\{z_1, \dots, z_n\}$ and we want to estimate the nonconformity of each example z_i in the bag, as in (2.15). (It is clear that the values (2.15) determine the nonconformity measure, and we will often define nonconformity measures by specifying the nonconformity scores (2.15).)

There is a natural solution if we are given a simple predictor (2.2) whose output does not depend on the order in which the old examples are presented. The simple predictor D then defines a prediction rule $D_{\{z_1, \dots, z_n\}} : \mathbf{X} \rightarrow \mathbf{Y}$ by the formula

$$D_{\{z_1, \dots, z_n\}}(x) := D(z_1, \dots, z_n, x).$$

A natural measure of nonconformity of z_i is the deviation of the predicted label

$$\hat{y}_i := D_{\{z_1, \dots, z_n\}}(x_i) \quad (2.21)$$

from the true label y_i . In this way any simple predictor, combined with a suitable measure of deviation of \hat{y}_i from y_i , leads to a nonconformity measure and, therefore, to a conformal predictor.

The simplest way of measuring the deviation of \hat{y}_i from y_i is to take the absolute value $|y_i - \hat{y}_i|$ of their difference as α_i . We could try, however, to somehow “standardize” $|y_i - \hat{y}_i|$ taking into account typical values we expect the difference between y_i and \hat{y}_i to take given the object x_i . Yet another approach is to take $\alpha_i := |y_i - \hat{y}_{(i)}|$, where $\hat{y}_{(i)}$ is the *deleted prediction*

$$\hat{y}_{(i)} := D_{\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}}(x_i)$$

computed by applying to x_i the prediction rule found from the data set with the example z_i deleted. The rationale behind this deletion is that z_i , even if it is an outlier, can influence the prediction rule $D_{\{z_1, \dots, z_n\}}$ so heavily that \hat{y}_i will become close to y_i , even though y_i can be very far from $\hat{y}_{(i)}$.

More generally, the prediction rule $D_{\{z_1, \dots, z_n\}}$ (or $D_{\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}}$) may map \mathbf{X} to some *prediction space* $\hat{\mathbf{Y}}$, not necessarily coinciding with \mathbf{Y} (e.g., $\hat{\mathbf{Y}}$ can be the set of all probability distributions on \mathbf{Y}). An *invariant simple predictor* is a function D that maps each bag $\{z_1, \dots, z_n\}$ of each size n to a prediction rule $D_{\{z_1, \dots, z_n\}} : \mathbf{X} \rightarrow \hat{\mathbf{Y}}$ and such that the function

$$(\{z_1, \dots, z_n\}, x) \mapsto D_{\{z_1, \dots, z_n\}}(x)$$

of the type $\mathbf{Z}^{(n)} \times \mathbf{X} \rightarrow \hat{\mathbf{Y}}$ is measurable for all n . A *discrepancy measure* is a measurable function $\Delta : \mathbf{Y} \times \hat{\mathbf{Y}} \rightarrow \mathbb{R}$. Given an invariant simple predictor D and a discrepancy measure Δ , we define functions A_n , $n = 1, 2, \dots$, as follows: for any $((x_1, y_1), \dots, (x_n, y_n)) \in \mathbf{Z}^*$, the values

$$\alpha_i = A_n((x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n) \setminus (x_i, y_i)) \quad (2.22)$$

are defined by the formula

$$\alpha_i := \Delta(y_i, D_{\{(x_1, y_1), \dots, (x_n, y_n)\} \setminus (x_i)}(x_i)) \quad (2.23)$$

or the formula

$$\alpha_i := \Delta(y_i, D_{\{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n)\}}(x_i)) . \quad (2.24)$$

It can be checked easily that in both cases (A_n) form a nonconformity measure.

2.3 Ridge regression confidence machine

In this section we will implement conformal prediction based on (2.23) and (2.24) concentrating on the case where the underlying simple predictor is ridge regression or nearest neighbors regression, which are two of the most standard regression algorithms. In the case of regression there is an obvious difficulty in implementing the idea of conformal prediction: it appears that to form a prediction set (2.17) we need to examine each potential classification y (cf. (2.19)). We will see, however, that there often is a feasible way to compute (2.17); in particular, this is the case for ridge regression and nearest neighbors regression. (We will make little effort to optimize the computational resources required, so “feasible” essentially means “avoiding examining infinitely many cases” here. Much faster algorithms will be constructed in §4.1.)

Least squares and ridge regression

Ridge regression and its special case, least squares, are among the most widely used regression algorithms. Least squares is the classical algorithm (going back to Gauss and Legendre), and ridge regression is its modification proposed in the 1960s. In this subsection we will describe least squares and ridge regression as simple predictors; for further details, see, e.g., Montgomery et al. 2001.

Suppose $\mathbf{X} = \mathbb{R}^p$ (objects are vectors consisting of p attributes), $\mathbf{Y} = \mathbb{R}$ (we are dealing with the problem of regression), and we are given a training set² z_1, \dots, z_n . To approximate the data, the ridge regression procedure recommends calculating the value $w \in \mathbb{R}^p$ where

$$a\|w\|^2 + \sum_{i=1}^n (y_i - w \cdot x_i)^2 \rightarrow \min \quad (2.25)$$

²It would be more correct to say “training sequence” or “training bag”, since we do not assume that all z_i are different, but we will use a more familiar term (as we already did in Chap. 1).

is attained; a is a nonnegative constant called the *ridge parameter*. The ridge regression prediction \hat{y} for the label y of an object x is then $\hat{y} := w \cdot x$. Least squares is the special case corresponding to $a = 0$.

We can naturally represent the ridge regression procedure in a matrix form. Let Y_n be the (column-) vector

$$Y_n := (y_1, \dots, y_n)' \quad (2.26)$$

of the labels and X_n be the $n \times p$ matrix formed from the objects

$$X_n := (x_1, x_2, \dots, x_n)' \quad (2.27)$$

Now we can represent the ridge regression procedure (2.25) as

$$a\|w\|^2 + \|Y_n - X_n w\|^2 \rightarrow \min , \quad (2.28)$$

or

$$Y_n' Y_n - 2w' X_n' Y_n + w' (X_n' X_n + aI_p) w \rightarrow \min .$$

Taking the derivative in w we obtain

$$2(X_n' X_n + aI_p)w - 2X_n' Y_n = 0 ,$$

or

$$w = (X_n' X_n + aI_p)^{-1} X_n' Y_n . \quad (2.29)$$

Standard statistical textbooks mainly discuss the case $a = 0$ (least squares). It is easy to see, however, that not only least squares is a special case of ridge regression, but ridge regression can be reduced to least squares as well: the solution of (2.28) for a general $a \geq 0$ can be found as the solution to the least squares problem

$$\|\bar{Y} - \bar{X}w\|^2 \rightarrow \min ,$$

where \bar{Y} is Y_n extended by adding p 0s on top and \bar{X} is X_n extended by adding the $p \times p$ matrix $\sqrt{a}I_p$ on top.

Basic RRCM

In this subsection we consider the conformal predictor determined by the nonconformity measure (2.22)–(2.23), with Δ the Euclidean distance and D the ridge regression procedure. Therefore, α_i are now the absolute values of the *residuals* $e_i := y_i - \hat{y}_i$, where \hat{y}_i is the ridge regression prediction for x_i based on the training set $x_1, y_1, \dots, x_n, y_n$. Two slightly more sophisticated approaches will be considered in the following subsection.

From (2.29) we can see that the ridge regression prediction for an object x is

$$x' w = x' (X_n' X_n + aI_p)^{-1} X_n' Y_n ; \quad (2.30)$$

therefore, the predictions \hat{y}_i for the objects x_i are given by

$$\hat{Y}_n := (\hat{y}_1, \dots, \hat{y}_n)' = X_n(X_n'X_n + aI_p)^{-1}X_n'Y_n.$$

The matrix

$$H_n := X_n(X_n'X_n + aI_p)^{-1}X_n' \quad (2.31)$$

is called the *hat matrix* (since it transforms the y_i into the hatted form \hat{y}_i) and plays an important role in the standard regression theory. This matrix, as well as $I_n - H_n$, is symmetric and idempotent when $a = 0$ (remember that a symmetric matrix M is *idempotent* if $MM = M$). Therefore, the vector of nonconformity scores $(\alpha_1, \dots, \alpha_n)'$ can be written in the form

$$|Y_n - H_n Y_n| = |(I_n - H_n)Y_n|.$$

Now suppose that we know the incomplete data sequence (2.4), we are given a significance level ϵ , and we want to compute the prediction set output by the conformal predictor determined by the nonconformity scores $\alpha_i = |e_i|$. Let y be a possible label for x_n and $Y := (y_1, \dots, y_{n-1}, y)'$. Note that $Y = (y_1, \dots, y_{n-1}, 0)' + (0, \dots, 0, y)'$ and so the vector of nonconformity scores can be represented as $|A + By|$ where

$$A = (I_n - H_n)(y_1, \dots, y_{n-1}, 0)'$$

and

$$B = (I_n - H_n)(0, \dots, 0, 1)'.$$

Therefore, each $\alpha_i = \alpha_i(y)$ varies piecewise-linearly as we change y . It is clear that the p-value $p(y)$ (defined to be the left-hand side of (2.18)) corresponding to y can only change at points where $\alpha_i(y) - \alpha_n(y)$ changes sign for some $i = 1, \dots, n-1$. This means that we can calculate the set of points y on the real line that have the p-value $p(y)$ exceeding ϵ rather than having to try all possible y , leading us to a feasible prediction algorithm.

For each $i = 1, \dots, n$, let

$$S_i := \{y : \alpha_i(y) \geq \alpha_n(y)\} = \{y : |a_i + b_i y| \geq |a_n + b_n y|\}, \quad (2.32)$$

where a_i and b_i are the components of A and B ($A = (a_1, \dots, a_n)'$ and $B = (b_1, \dots, b_n)'$). Each set S_i (always closed) will either be the real line, the union of two rays, a ray, an interval, a point (which is a special case of an interval), or empty. Indeed, as we are interested in $|a_i + b_i y|$ we can assume $b_i \geq 0$ for $i = 1, \dots, n$ (if necessary, multiply both a_i and b_i by -1). If $b_i \neq b_n$ then $\alpha_i(y)$ and $\alpha_n(y)$ are equal at two points (which may coincide):

$$-\frac{a_i - a_n}{b_i - b_n} \quad \text{and} \quad -\frac{a_i + a_n}{b_i + b_n}; \quad (2.33)$$

in this case, S_i is an interval (possibly a point) or the union of two rays. If $b_i = b_n \neq 0$ then $\alpha_i(y) = \alpha_n(y)$ at the only point

$$-\frac{a_i + a_n}{2b_i} \quad (2.34)$$

(and S_i is a ray) unless $a_i = a_n$, in which case $S_i = \mathbb{R}$. If $b_i = b_n = 0$, S_i is either \emptyset or \mathbb{R} .

To calculate the p-value $p(y)$ for any potential label y of x_n , we count how many S_i include y and divide by n :

$$p(y) = \frac{|\{i = 1, \dots, n : y \in S_i\}|}{n}.$$

It is clear that as y increases $p(y)$ can only change at the points (2.33) and (2.34) and so for any threshold ϵ we can find the explicit representation of the set of y for which $p(y) > \epsilon$ as the union of finitely many intervals and rays. The following algorithm gives a slightly easier to describe implementation of this idea; it arranges the points (2.33) and (2.34) into an increasing sequence $y_{(1)}, \dots, y_{(m)}$, adds $y_{(0)} := -\infty$ and $y_{(m+1)} := \infty$, and then computes $N(j)$, the number of i such that $(y_{(j)}, y_{(j+1)}) \subseteq S_i$, for $j = 0, \dots, m$, and $M(j)$, the number of i such that $y_{(j)} \in S_i$, for $j = 1, \dots, m$. The algorithm is given a (small) set of significance levels ϵ_k , $k = 1, \dots, K$, and outputs the corresponding nested family of prediction sets $\Gamma_n^{\epsilon_k}$, $k = 1, \dots, K$.

ALGORITHM RRCM (RIDGE REGRESSION CONFIDENCE MACHINE)

```

 $C := I_n - X_n(X_n'X_n + aI_p)^{-1}X_n'$ ,  $X_n$  being defined by (2.27);
 $A = (a_1, \dots, a_n)' := C(y_1, \dots, y_{n-1}, 0)'$ ;
 $B = (b_1, \dots, b_n)' := C(0, \dots, 0, 1)'$ ;
FOR  $i = 1, \dots, n$ :
  IF  $b_i < 0$  THEN  $a_i := -a_i$ ;  $b_i := -b_i$  END IF
END FOR;
 $P := \emptyset$ ;
FOR  $i = 1, \dots, n$ :
  IF  $b_i \neq b_n$  THEN add (2.33) to  $P$  END IF;
  IF  $b_i = b_n \neq 0$  AND  $a_i \neq a_n$  THEN add (2.34) to  $P$  END IF
END FOR;
sort  $P$  in ascending order obtaining  $y_{(1)}, \dots, y_{(m)}$ ;
set  $y_{(0)} := -\infty$  and  $y_{(m+1)} := \infty$ ;
 $N(j) := 0$ ,  $j = 0, \dots, m$ ;
FOR  $i = 1, \dots, n$ :
  FOR  $j = 0, \dots, m$ :
    IF  $|a_i + b_i y| \geq |a_n + b_n y|$  for  $y \in (y_{(j)}, y_{(j+1)})$ 
    THEN  $N(j) := N(j) + 1$ 
    END IF
  END FOR
END FOR;
 $M(j) := 0$ ,  $j = 1, \dots, m$ ;
FOR  $i = 1, \dots, n$ :
```

```

FOR  $j = 1, \dots, m$ :
  IF  $|a_i + b_i y_{(j)}| \geq |a_n + b_n y_{(j)}|$ 
    THEN  $M(j) := M(j) + 1$ 
  END IF
END FOR
END FOR;
FOR  $k = 1, \dots, K$ :
   $\Gamma_n^{\epsilon_k} := (\cup_{j:N(j)/n > \epsilon_k} (y_{(j)}, y_{(j+1)})) \cup \{y_{(j)} : M(j)/n > \epsilon_k\}$ 
END FOR.

```

This algorithm is run by Predictor at each trial $n = 1, 2, \dots$ of the on-line prediction protocol (given on p. 20).

Let us suppose that the number p of attributes is constant. It is clear that Algorithm RRCM requires computation time $O(n^2)$. A simple modification of the algorithm, however, reduces the required computation time to $O(n \log n)$.

First, it is clear that

$$A = (y_1, \dots, y_{n-1}, 0)' - X_n [(X_n' X_n + a I_p)^{-1} X_n' (y_1, \dots, y_{n-1}, 0)']$$

and

$$B = (0, \dots, 0, 1)' - X_n [(X_n' X_n + a I_p)^{-1} X_n' (0, \dots, 0, 1)']$$

can be computed in time $O(n)$. Sorting P can be done in time $O(n \log n)$ (see, e.g., Cormen et al. 2001, Part II). Therefore, it suffices to show that the two double loops (computing N and computing M) in Algorithm RRCM can be implemented in time $O(n)$. Instead of computing the array $N(j)$, $j = 0, \dots, m$, directly, we can first compute $N'(j) := N(j) - N(j-1)$, $j = 0, \dots, m$, with $N(-1) := 0$; it is easy (takes time $O(n)$) to compute N from N' . Analogously, we can compute $M'(j) := M(j) - M(j-1)$, $j = 1, \dots, m$, with $M(0) := 0$, instead of M . To find N' and M' in time $O(n)$, initialize $N'(j) := 0$, $j = 0, \dots, m$, $M'(j) := 0$, $j = 1, \dots, m$, and for each example $i = 1, \dots, n$ do the following:

- if S_i (see (2.32)) is empty, do nothing;
- if S_i contains only one point, $S_i = \{y_{(j)}\}$, set $M'(j) := M'(j) + 1$ and $M'(j+1) := M'(j+1) - 1$ (assignments that do not make sense, such as $M'(j+1) := M'(j+1) - 1$ for $j = m$, are simply ignored);
- if S_i is an interval $[y_{(j_1)}, y_{(j_2)}]$ and $j_1 < j_2$, set $M'(j_1) := M'(j_1) + 1$, $M'(j_2+1) := M'(j_2+1) - 1$, $N'(j_1) := N'(j_1) + 1$, $N'(j_2) := N'(j_2) - 1$;
- if S_i is a ray $(-\infty, y_{(j)})$, set $M'(1) := M'(1) + 1$, $M'(j+1) := M'(j+1) - 1$, $N'(0) := N'(0) + 1$, $N'(j) := N'(j) - 1$;
- if S_i is a ray $[y_{(j)}, \infty)$, set $M'(j) := M'(j) + 1$, $N'(j) := N'(j) + 1$;
- if S_i is the union $(-\infty, y_{(j_1)}) \cup [y_{(j_2)}, \infty)$ of two rays such that $j_1 < j_2$, set $M'(1) := M'(1) + 1$, $M'(j_1+1) := M'(j_1+1) - 1$, $M'(j_2) := M'(j_2) + 1$, $N'(0) := N'(0) + 1$, $N'(j_1) := N'(j_1) - 1$, $N'(j_2) := N'(j_2) + 1$;

- if S_i is the real line $(-\infty, \infty)$, set $M'(1) := M'(1) + 1$ and $N'(0) := N'(0) + 1$.

Two modifications

The algorithm developed in the previous subsection can be easily modified to allow two alternative ways of computing nonconformity scores. To simplify formulas, we assume $a = 0$ in this subsection (i.e., we will consider least squares). A *least squares confidence machine* (LSCM) is an RRCM with the ridge parameter a set to 0.

First we consider the special case of (2.24) where α_i is defined to be the absolute value $|y_i - \hat{y}_{(i)}|$ of the *deleted residual* $e_{(i)} := y_i - \hat{y}_{(-i)}$, where $\hat{y}_{(-i)}$ is the least squares prediction for y_i computed from x_i based on the training set $x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_{i+1}, y_{i+1}, \dots, x_n, y_n$. It is well known in statistics that to compute the deleted residuals $e_{(i)}$ we do not need to perform n regressions; they can be easily computed from the usual residuals $e_i = |y_i - \hat{y}_i|$ by the formula

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} , \quad (2.35)$$

where h_{ii} is the i th diagonal element of the hat matrix H . (For a proof, see Montgomery et al. 2001, Appendix C.7.)

Let us call the conformal predictor determined by the nonconformity scores $\alpha_i := |e_{(i)}|$ the *deleted LSCM*. Algorithm LSCM will implement the deleted LSCM if A and B are redefined as follows:

$$a_i := \frac{a_i}{1 - h_{ii}}, \quad b_i := \frac{b_i}{1 - h_{ii}}, \quad i = 1, \dots, n .$$

It is clear that each

$$h_{ii} = x_i' (X_n' X_n)^{-1} x_i$$

can be computed from $(X_n' X_n)^{-1}$ in time $O(1)$ (again assuming that the number p of attributes is constant), and so the deleted LSCM can also be implemented in time $O(n \log n)$.

Another natural modification of LSCM is half-way between the LSCM and the deleted LSCM: the nonconformity scores are taken to be

$$\alpha_i := \frac{|e_i|}{\sqrt{1 - h_{ii}}} . \quad (2.36)$$

We will explain the motivation behind this choice momentarily, but first describe how to implement the *studentized LSCM* determined by these nonconformity scores. The implementation is just Algorithm LSCM with A and B redefined as

$$a_i := \frac{a_i}{\sqrt{1 - h_{ii}}}, \quad b_i := \frac{b_i}{\sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n .$$

The computation time is again $O(n \log n)$.

Now we explain the standard motivation for studentized LSCM. (Remember that this motivation has no bearing on the validity of the constructed conformal predictor; in particular, the smoothed version of this confidence predictor will make errors independently with probability ϵ at each significance level ϵ regardless whether the assumption of normal noise we are about to make is satisfied or not; cf. §10.3.) Imagine that the labels y_i are generated from the deterministic objects x_i in the following way:

$$y_i = w \cdot x_i + \xi_i , \quad (2.37)$$

where ξ_i are independent normal random variables with the mean 0 and same variance σ^2 (random noise). Set $\xi := (\xi_1, \dots, \xi_n)'$. Since the vector of residuals is $e = (I_n - H_n)Y_n$ (see above), we obtain

$$e = (I_n - H_n)(X_n w + \xi) = (I_n - H_n)\xi \quad (2.38)$$

for any fixed w (the true parameters); therefore, the covariance matrix of the residuals is

$$\text{var}(e) = \text{var}((I_n - H_n)\xi) = (I_n - H_n)\text{var}(\xi)(I_n - H_n)' = \sigma^2(I_n - H_n) ,$$

since $\text{var}(\xi) = \sigma^2 I_n$ and $I_n - H_n$ is symmetric and idempotent. We can see that the variance of e_i is $(1 - h_{ii})\sigma^2$, and the scaling of the residuals e_i by dividing by $\sqrt{1 - h_{ii}}$ will equalize their variances (and even their distributions, since by (2.38), e_i are normally distributed).

Notice that according to our motivational model (2.37) the level of noise ξ_i does not depend on the observed object x_i (the variance of ξ_i remains the same, σ^2). Even in this case, it may be useful to scale residuals. If we suspect that noise can be different in different parts of the object space, heavier scaling may become necessary for satisfactory prediction.

Dual form ridge regression

Least squares and ridge regression procedures can only deal with situations where the number of parameters p is relatively small since they involve inverting a $p \times p$ matrix. They are carried over to high-dimensional problems using the so-called “kernel trick”, introduced in machine learning by Vapnik (see, e.g., Vapnik 1995, 1998). (The formulas arrived at by using the kernel trick coincide with those obtained by means of Gaussian processes and reproducing kernel Hilbert spaces; see §2.7.)

We first state the ridge regression procedure in the “dual form”. The traditional statistical approach to dualization is to use the easy-to-check matrix equality

$$X_n(X'_n X_n + aI_p)^{-1} = (X_n X'_n + aI_n)^{-1} X_n , \quad (2.39)$$

which can be equivalently rewritten as

$$(X_n' X_n + aI_p)^{-1} X_n' = X_n' (X_n X_n' + aI_n)^{-1}.$$

(To see that (2.39) is true, multiply it by $(X_n X_n' + aI_n)$ on the left and $(X_n' X_n + aI_p)$ on the right.) Using (2.39), we can rewrite the ridge regression prediction for an object x based on examples $(x_1, y_1, \dots, x_n, y_n)$ as

$$\hat{y} = Y_n' X_n (X_n' X_n + aI_p)^{-1} x = Y_n' (X_n X_n' + aI_n)^{-1} X_n x, \quad (2.40)$$

where the $n \times p$ matrix X_n and vector Y_n are defined as before. The crucial property of the representation $\hat{y} = Y_n' (X_n X_n' + aI_n)^{-1} X_n x$ is that it depends on the objects x_1, \dots, x_n, x only via the scalar products between them. In particular, if the object space \mathbf{X} is mapped into another Euclidean space (called the *feature space*) \mathbf{H} , $F : \mathbf{X} \rightarrow \mathbf{H}$, and ridge regression is performed in the feature space, the prediction (2.40) can be written in the form

$$\hat{y} = Y_n' (K_n + aI_n)^{-1} k_n, \quad (2.41)$$

where K_n is the matrix with elements $(K_n)_{i,j} := \mathcal{K}(x_i, x_j)$, k_n is the vector with elements $(k_n)_i := \mathcal{K}(x, x_i)$, and \mathcal{K} is the *kernel*, defined by

$$\mathcal{K}(x^{(1)}, x^{(2)}) := F(x^{(1)}) \cdot F(x^{(2)}) \quad (2.42)$$

for all $x^{(1)}, x^{(2)} \in \mathbf{X}$. The hat matrix in the dual representation is

$$H_n = X_n (X_n' X_n + aI_p)^{-1} X_n' = (X_n X_n' + aI_n)^{-1} X_n X_n',$$

and if ridge regression is carried out in the feature space, this becomes

$$H_n = (K_n + aI_n)^{-1} K_n. \quad (2.43)$$

Now it is easy to represent the RRCM algorithm in the kernel form: the only difference from Algorithm RRCM is that now C is defined as $I_n - H_n$ with H_n given by (2.43).

The computation time of the kernel form of the RRCM algorithm is $O(n^2)$. This can be seen from the well-known (see, e.g., Henderson and Searle 1981, (8)) and easy-to-check formula

$$\begin{pmatrix} K & k \\ k' & \kappa \end{pmatrix}^{-1} = \begin{pmatrix} K^{-1} + dK^{-1}kk'K^{-1} & -dK^{-1}k \\ -dk'K^{-1} & d \end{pmatrix}, \quad (2.44)$$

where K is a square matrix, k a vector, κ a number, and

$$d := \frac{1}{\kappa - k' K^{-1} k}.$$

Indeed, by this formula $(K_n + aI_n)^{-1}$ can be updated from the previous trial of the on-line learning protocol in time $O(n^2)$, and both

$$A = (y_1, \dots, y_{n-1}, 0)' - (K_n + aI_n)^{-1} (K_n(y_1, \dots, y_{n-1}, 0)')$$

and

$$B = (0, \dots, 0, 1)' - (K_n + aI_n)^{-1} (K_n(0, \dots, 0, 1)')$$

can be computed in time $O(n^2)$. There are some conditions on the validity of formula (2.44), but they are satisfied in our context: the theorem on normal correlation (see, e.g., Shiryaev 1996, Theorem II.13.2) implies that d is well-defined (and positive) whenever the matrix $\begin{pmatrix} K & k \\ k' & \kappa \end{pmatrix}$ is positive definite; the latter condition is satisfied when the formula is used for updating $(K_n + aI_n)^{-1}$, $a > 0$, to $(K_{n+1} + aI_{n+1})^{-1}$ (in which case $K = K_n + aI_n$, $k = k_n$, and $\kappa = \mathcal{K}(x_n, x_n) + a$).

It is easy to see that the above construction also works in the case where \mathbf{H} is an arbitrary Hilbert space. The essence of the kernel trick is that one does not need to consider the feature space in explicit form (Vapnik 1998, §10.5.2). It is clear that any kernel (2.42) is symmetric,

$$\mathcal{K}(x^{(1)}, x^{(2)}) = \mathcal{K}(x^{(2)}, x^{(1)}), \quad \forall x^{(1)}, x^{(2)} \in \mathbf{X}, \quad (2.45)$$

and nonnegative definite,

$$\sum_{i=1}^m \sum_{j=1}^m \mathcal{K}(x^{(i)}, x^{(j)}) a_i a_j \geq 0, \quad \forall x^{(1)}, \dots, x^{(m)} \in \mathbf{X}, \forall a_1, \dots, a_m \in \mathbb{R}. \quad (2.46)$$

(To see that (2.46) is true, notice that

$$\begin{aligned} & \|a_1 F(x^{(1)}) + \dots + a_m F(x^{(m)})\|^2 \\ &= \left(a_1 F(x^{(1)}) + \dots + a_m F(x^{(m)}) \right) \cdot \left(a_1 F(x^{(1)}) + \dots + a_m F(x^{(m)}) \right) \geq 0. \end{aligned}$$

It turns out that the opposite statement is also true: any function $\mathcal{K} : \mathbf{X}^2 \rightarrow \mathbb{R}$ that is symmetric and nonnegative definite can be represented in the form (2.42). There are many proofs of this result, but one of the simplest arguments is “probabilistic”: any symmetric nonnegative definite matrix is the covariance matrix of a set of (zero-mean) normal random variables; the Daniell–Kolmogorov (Kolmogorov 1933a) theorem then immediately implies that any symmetric nonnegative definite function $\mathcal{K}(x^{(1)}, x^{(2)})$ is the “infinite covariance matrix” of a zero-mean Gaussian random field $\xi(x)$, $x \in \mathbf{X}$:

$$\mathcal{K}(x^{(1)}, x^{(2)}) = \mathbb{E} \left(\xi(x^{(1)}) \xi(x^{(2)}) \right).$$

The last equality is a special case of (2.42) since the zero-mean finite-variance random variables with dot product

$$\xi \cdot \eta := \mathbb{E}(\xi \eta)$$

form a Hilbert space (not necessarily separable; see, e.g., Shiryaev 1996, §III.11).

Nearest neighbors regression

Least squares and ridge regression are just two of the standard regression algorithms; conformal predictors can be implemented in a feasible way for nonconformity measures based on many other regression algorithms. Such an implementation is especially simple in the case of the nearest neighbors algorithm.

The idea of the k -nearest neighbors algorithms, where k is the number of “neighbors” taken into account, is as follows. Suppose the object space \mathbf{X} is a metric space (for example, the usual Euclidean distance is often used if $\mathbf{X} = \mathbb{R}^p$). To give a prediction for a new object x_n , find the k objects x_{i_1}, \dots, x_{i_k} among the known examples that are nearest to x_n in the sense of the chosen metric (assuming, for simplicity, that there are no ties). In the problem of classification, the predicted classification \hat{y}_n is obtained by “voting”: it is defined to be the most frequent label among y_{i_1}, \dots, y_{i_k} . In regression, we can take, e.g., the mean or the median of y_{i_1}, \dots, y_{i_k} .

We will only consider the version of the k -nearest neighbors regression (k -NNR) where the prediction \hat{y} for a new object x based on the training set (x_i, y_i) , $i = 1, \dots, n$, is defined to be the arithmetic mean of the labels of the k nearest neighbors of x among x_1, \dots, x_n . It will be easy to see that the more robust procedure where arithmetic mean is replaced by median also leads to a feasible conformal predictor.

Consider the special case of the nonconformity scores (2.24) where $\alpha_i := |y_i - \hat{y}_{(i)}|$ and $\hat{y}_{(i)}$ is the k -NNR prediction for x_i based on the training set (x_j, y_j) , $j = 1, \dots, i-1, i+1, \dots, n$. The conformal predictor determined by this nonconformity measure (*k -NNR conformal predictor*) is implemented by the RRCM algorithm with the only modification that a_i and b_i are now defined as follows (we assume that $n > k$ and that all distances between the objects are different):

- a_n is the minus arithmetic mean of the labels of x_n ’s k nearest neighbors and $b_n = 1$;
- if $i < n$ and x_n is among the k nearest neighbors of x_i , a_i is x_i ’s label minus the arithmetic mean of the labels of those nearest neighbors with x_n ’s label set to 0, and $b_i = -1/k$;
- if $i < n$ and x_n is not among the k nearest neighbors of x_i , a_i is x_i ’s label minus the arithmetic mean of the labels of x_i ’s k nearest neighbors, and $b_i = 0$.

(It is obvious that the nonconformity score for the i th example is $\alpha_i = |a_i + b_i y|$, where y is the label for x_n that is being tried.)

Instead of measuring distance in the original example space \mathbf{X} we can measure it in the feature space, which corresponds to using the function

$$\begin{aligned}
\rho(x^{(1)}, x^{(2)}) &:= \|F(x^{(1)}) - F(x^{(2)})\|^2 \\
&= \left(F(x^{(1)}) - F(x^{(2)})\right) \cdot \left(F(x^{(1)}) - F(x^{(2)})\right) \\
&= \mathcal{K}(x^{(1)}, x^{(1)}) + \mathcal{K}(x^{(2)}, x^{(2)}) - 2\mathcal{K}(x^{(1)}, x^{(2)})
\end{aligned}$$

as the distance (remember that in the case of k -NNR conformal prediction we are only interested in the distance up to a monotonic transformation).

Experimental results

In this subsection we empirically test the validity and evaluate the efficiency of some of the conformal predictors introduced earlier. We start from RRCM, reporting only results for $a = 1$ on the randomly permuted Boston Housing data set (this data set and the rationale for shuffling, counteracting possible deviations from exchangeability, are described in Appendix B). A dummy attribute always taking value 1 (to allow a non-zero intercept) was added to each example, and at each trial each attribute was linearly scaled for the known objects to span the interval $[-1, 1]$ (or $[0, 0]$, if the attribute took the same value for all known objects, as described in Appendix B). Figures 2.1–2.3 show the performance of RRCM in regard of its efficiency. In Fig. 2.1, the solid line shows, for each $n = 1, \dots, 506$, the median $M_n^{99\%}$ of the widths of the convex hulls $\text{co } \Gamma_i^{1\%}$ of the prediction sets $\Gamma_i^{1\%}$, $i = 1, \dots, n$, at confidence level 99%; similarly, the dashed line shows $M_n^{95\%}$ and the dash-dot line shows $M_n^{80\%}$. Figure 2.2 presents more detailed information for the performance at the confidence level 95%: not only the median $M_n^{95\%}$ of the widths of $\text{co } \Gamma_i^{5\%}$, $i = 1, \dots, n$, but also the upper and lower quartiles of those widths; the cumulative number of errors at this confidence level is also given. Both Figs. 2.1 and 2.2 give, for comparison, a graph reflecting the performance of the ridge regression procedure as a simple predictor: the dotted line shows the sequence $2A(n)$, where $A(n)$ is the median of the sequence $|y_i - \hat{y}_i|$, $i = 1, \dots, n$, of distances between the true label y_i and the prediction \hat{y}_i given by the ridge regression procedure according to (2.30). This line lies well below the other lines, the main reason being that we considered confidence levels of 80% and above; lowering the plank to 50% leads to near coincidence (Fig. 2.3). The cumulative error lines are close to straight lines with correct slopes (Fig. 2.4), although because of the small sample size the imperfections due to statistical fluctuations are more noticeable than in Fig. 1.5 on p. 10.

In fact, the RRCM's performance was not sensitive to moderate changes in a : e.g., running the algorithm for $a = 0$ (LSCM), whether modified (deleted or studentized LSCM) or not, produced virtually identical figures.

The quality of prediction can be improved by using non-linear methods. Figure 2.5 shows the performance of the kernel RRCM with the second-order polynomial kernel

$$\mathcal{K}(x^{(1)}, x^{(2)}) := \left(1 + x^{(1)} \cdot x^{(2)}\right)^2$$

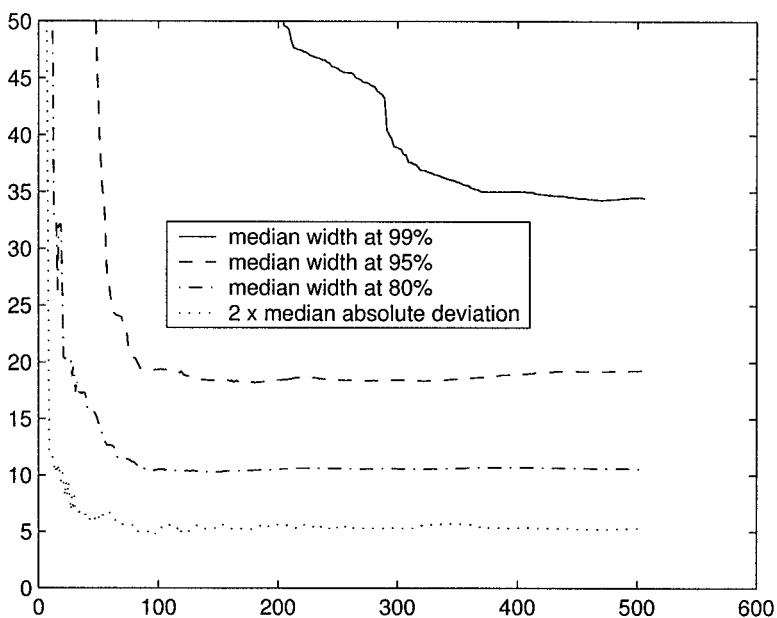


Fig. 2.1. The on-line performance of RRCM on the randomly permuted Boston Housing data set (of size 506)

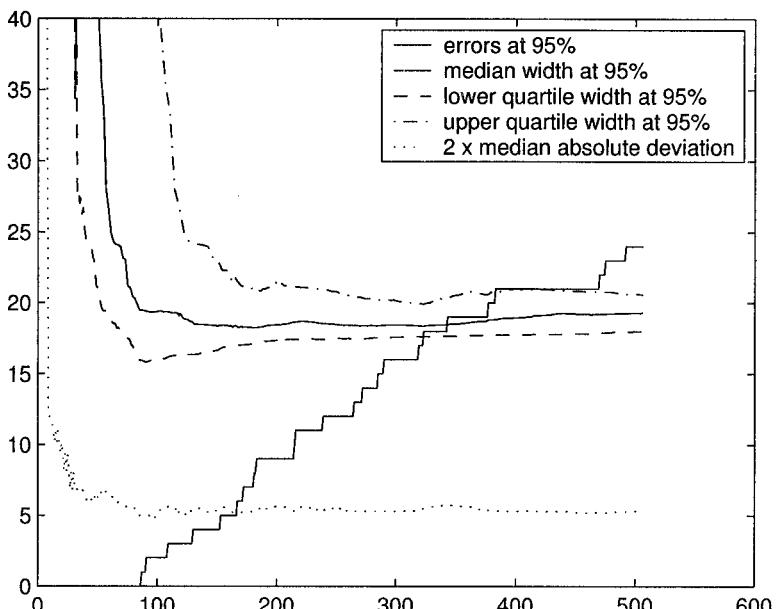


Fig. 2.2. The on-line performance of RRCM on the randomly permuted Boston Housing data set at the confidence level 95%

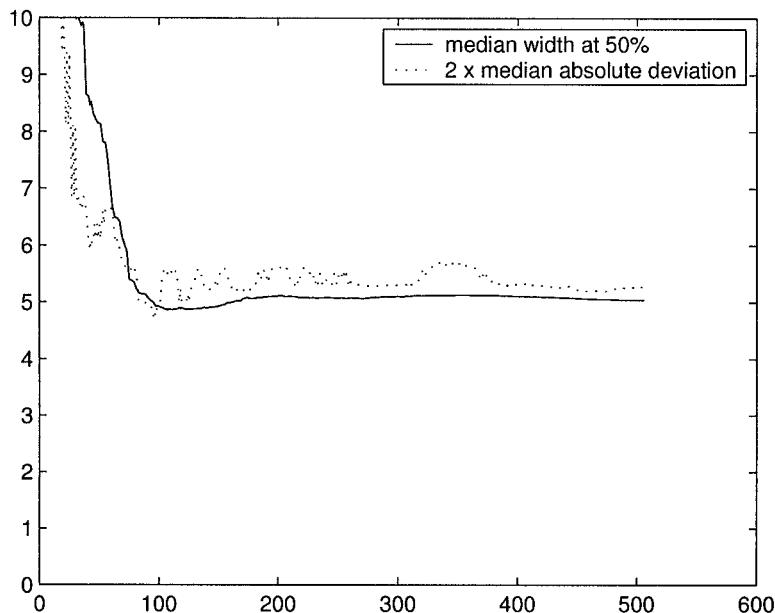


Fig. 2.3. The on-line performance of RRCM on the randomly permuted Boston Housing data set at the confidence level 50%

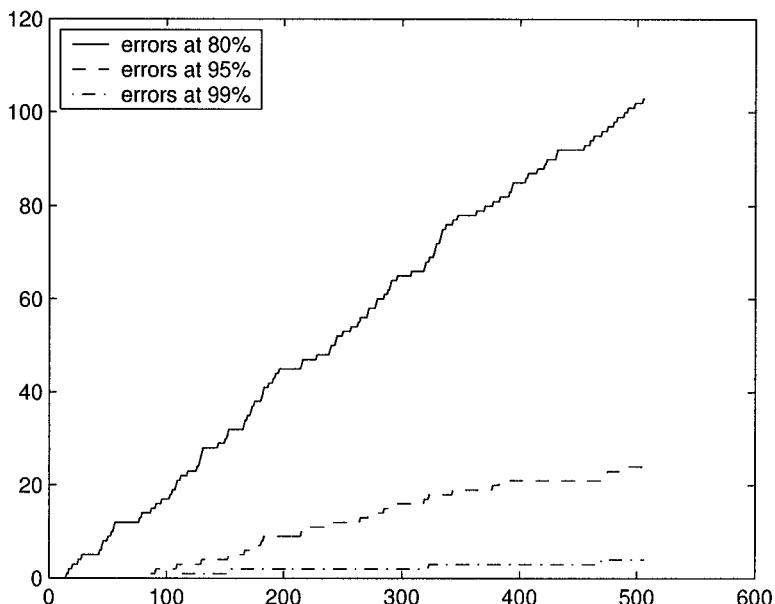


Fig. 2.4. The cumulative numbers of errors at the given confidence levels for RRCM run on-line on the randomly permuted Boston Housing data set

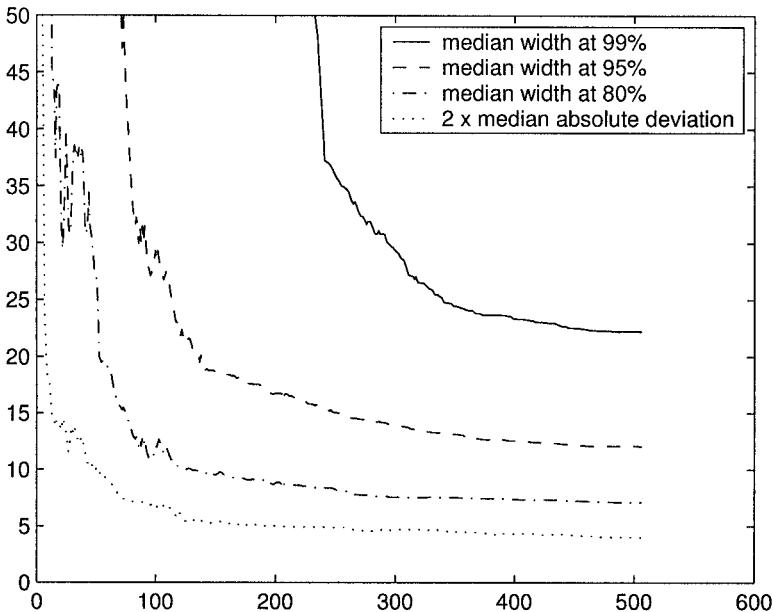


Fig. 2.5. The on-line performance of kernel RRCM on the randomly permuted Boston Housing data set

using the same format as Fig. 2.1.

The performance of the 1-NNR conformal predictor (as described in the preceding subsection) is shown, in the same format, in Fig. 2.6. The 1-NNR procedure performs reasonably well as a simple predictor (as the dotted line shows), but the prediction intervals it produces are much worse than those produced by more advanced methods.

2.4 Are there other ways to achieve validity?

In this section we will see that conformal predictors are essentially the only confidence predictors in a very natural class that satisfy our strong non-asymptotic property of validity.

Let us say that a confidence predictor is *invariant* if $\Gamma^\epsilon(z_1, \dots, z_{n-1}, x_n)$ does not depend on the order in which z_1, \dots, z_{n-1} are listed. Since we assume exchangeability, the invariant confidence predictors constitute a natural class (see, e.g., the description of the “sufficiency principle” in Cox and Hinkley 1974; later in this book, however, we will also study confidence predictors that are not invariant, such as inductive and Mondrian conformal predictors in Chap. 4).

If Γ_1 and Γ_2 are (deterministic) confidence predictors, we will say that Γ_1 is *at least as good as* Γ_2 if, for any n and any $\epsilon \in (0, 1)$,

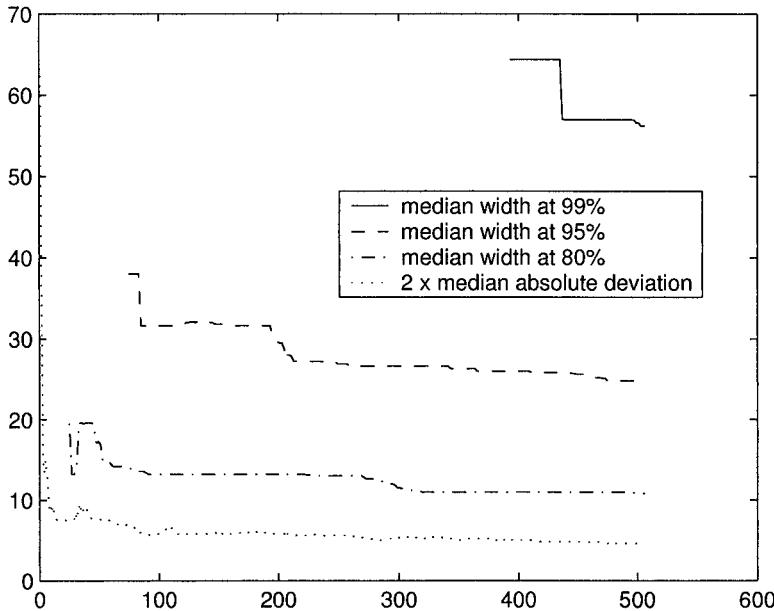


Fig. 2.6. The on-line performance of the 1-NNR conformal predictor on the randomly permuted Boston Housing data set

$$\Gamma_1^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma_2^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$$

holds for almost all $x_1, y_1, x_2, y_2, \dots$ generated by any exchangeable distribution on \mathbf{Z}^∞ .

The following proposition asserts that invariant conservatively valid confidence predictors are conformal predictors or can be improved to become conformal predictors.

Theorem 2.6. *Suppose \mathbf{Z} is a Borel space. Let Γ be an invariant conservatively valid confidence predictor. Then there is a conformal predictor that is at least as good as Γ .*

This proposition will be proved in §2.6 (p. 48). It is interesting that the proof will not use the fact that the random variables $\xi_1^{(\epsilon)}, \xi_2^{(\epsilon)}, \dots$ from the definition of Γ 's conservative validity are independent. This observation leads to the following simple result, which we state in terms of randomized confidence predictors.

Proposition 2.7. *Suppose \mathbf{Z} is Borel. If an invariant randomized confidence predictor Γ at each significance level $\epsilon \in (0, 1)$ makes an error with probability ϵ at each trial and under any exchangeable distribution on \mathbf{Z}^∞ , then it makes errors at different trials independently, at each significance level $\epsilon \in (0, 1)$ and under any exchangeable distribution on \mathbf{Z}^∞ .*

2.5 Conformal transducers

There are two convenient ways to represent a conformal predictor: as a confidence predictor and as a “transducer”. So far we have been using the first way; the goal of this section is to introduce the second way, which is simpler mathematically, and to discuss connections between the two. This will be needed in, e.g., Chap. 7, and can be skipped for now.

A *randomized transducer* is a function f of the type $(\mathbf{X} \times [0, 1] \times \mathbf{Y})^* \rightarrow [0, 1]$. It is called “transducer” because it can be regarded as mapping each input sequence $(x_1, \tau_1, y_1, x_2, \tau_2, y_2, \dots)$ in $(\mathbf{X} \times [0, 1] \times \mathbf{Y})^\infty$ into the output sequence (p_1, p_2, \dots) of *p-values* defined by $p_n := f(x_1, \tau_1, y_1, \dots, x_n, \tau_n, y_n)$, $n = 1, 2, \dots$. We say that f is an *exactly valid randomized transducer* (or just *exact randomized transducer*) if the output p-values p_1, p_2, \dots are always distributed according to the uniform distribution \mathbf{U}^∞ on $[0, 1]^\infty$, provided the input examples $z_n = (x_n, y_n)$, $n = 1, 2, \dots$, are generated by an exchangeable probability distribution on \mathbf{Z}^∞ and the numbers τ_1, τ_2, \dots are generated independently from the uniform distribution \mathbf{U} on $[0, 1]$.

We can extract exact randomized transducers from nonconformity measures: given a nonconformity measure A , for each sequence

$$(x_1, \tau_1, y_1, \dots, x_n, \tau_n, y_n) \in (\mathbf{X} \times [0, 1] \times \mathbf{Y})^*$$

define

$$f(x_1, \tau_1, y_1, \dots, x_n, \tau_n, y_n) := \frac{|\{i : \alpha_i > \alpha_n\}| + \tau_n |\{i : \alpha_i = \alpha_n\}|}{n}, \quad (2.47)$$

where α_i , $i = 1, 2, \dots$, are computed from $z_i = (x_i, y_i)$ using A by the usual (cf. (2.19), p. 26) formula

$$\alpha_i := A_n(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}, z_i).$$

Each randomized transducer f that can be obtained in this way will be called a *smoothed conformal transducer*.

Proposition 2.8. *Each smoothed conformal transducer is an exact randomized transducer.*

This proposition is a special case of Theorem 8.1 (p. 193), which will be proved in §8.7.

In a similar way we can define (deterministic) *conformal transducers* f : given a nonconformity measure A , for each sequence $(z_1, \dots, z_n) \in \mathbf{Z}^*$ set

$$f(z_1, \dots, z_n) := \frac{|\{i : \alpha_i \geq \alpha_n\}|}{n},$$

where α_i are computed as before. In general, a (deterministic) *transducer* is a function f of the type $\mathbf{Z}^* \rightarrow [0, 1]$; as before, we associate with f a mapping

from $z_1 z_2 \dots$ to the *p-values* $p_1 p_2 \dots$ ($p_n := f(z_1, \dots, z_n)$). We say that f is a *conservatively valid transducer* (or *conservative transducer*) if there exists a probability space with two sequences ξ_n and η_n , $n = 1, 2, \dots$, of $[0, 1]$ -valued random variables such that:

- the sequence ξ_1, ξ_2, \dots is distributed as \mathbf{U}^∞ ;
- for each n , $\eta_n \leq \xi_n$;
- the joint distribution of the sequence of p-values produced by f coincides with the joint distribution of η_1, η_2, \dots , provided the examples z_1, z_2, \dots are generated from an exchangeable distribution on \mathbf{Z}^∞ .

The following implication of Proposition 2.8 is obvious:

Corollary 2.9. *Each conformal transducer is conservative.*

We can fruitfully discuss confidence transducers even in the case of general example spaces \mathbf{Z} , not necessarily products $\mathbf{X} \times \mathbf{Y}$ of object and label spaces. But in the latter case we can associate a confidence predictor $\Gamma = f'$ with each confidence transducer f defining (2.17) (p. 26) as

$$\{y \in \mathbf{Y} : f(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, y) > \epsilon\}. \quad (2.48)$$

Vice versa, with any confidence predictor Γ we can associate the confidence transducer $f = \Gamma'$ defined by

$$f(x_1, y_1, \dots, x_n, y_n) := \sup \{\epsilon : y_n \in \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)\}. \quad (2.49)$$

Letting x_i in (2.48) and (2.49) range over the extended object space $\mathbf{X} \times [0, 1]$, we obtain the definition of the randomized confidence predictor f' associated with a randomized confidence transducer f and the definition of the randomized confidence transducer Γ' associated with a randomized confidence predictor Γ . The definition (2.49) of p-values associated with a confidence transducer agrees with the definitions given earlier (see (2.16) and the left-hand sides of (2.18) and (2.20)).

We will see in the next subsection that the expositions of the theory of hedged prediction in terms of conformal transducers and conformal predictors are essentially equivalent. But first we slightly strengthen Proposition 2.4.

A randomized confidence predictor is *strongly exact* if, for any exchangeable probability distribution on \mathbf{Z}^∞ and any sequence $(\epsilon_1, \epsilon_2, \dots) \in (0, 1)^\infty$ of significance levels, the sequence of random variables $\text{err}_n^{\epsilon_n}(\Gamma, P)$, $n = 1, 2, \dots$, is distributed as the product $\mathbf{B}_{\epsilon_1} \times \mathbf{B}_{\epsilon_2} \times \dots$ of Bernoulli distributions with parameters $\epsilon_1, \epsilon_2, \dots$. It can be defined in a similar way what it means for a confidence predictor Γ to be strongly conservative.

Theorem 8.1 will also imply the following proposition.

Proposition 2.10. *Any smoothed conformal predictor is strongly exact.*

Normalized confidence predictors and confidence transducers

To obtain full equivalence between confidence transducers and confidence predictors, a further natural restriction has to be imposed on the latter: they will be required to be “normalized”. This is a mild restriction since each confidence predictor can be normalized in such a way that its quality does not suffer.

Formally, the *normal form* Γ_{norm} of a confidence predictor Γ is defined by

$$\Gamma_{\text{norm}}^{\epsilon}(x_1, y_1, \dots, x_n) := \bigcup_{\epsilon' > \epsilon} \Gamma^{\epsilon'}(x_1, y_1, \dots, x_n). \quad (2.50)$$

We say that Γ is *normalized* if $\Gamma_{\text{norm}} = \Gamma$. These definitions are also applicable to randomized confidence predictors, in which case x_i range over the extended object space $\mathbf{X} \times [0, 1]$. The following proposition lists some basic properties of the operation norm and normalized confidence predictors.

Proposition 2.11. *All conformal predictors and smoothed conformal predictors are normalized. For any confidence predictor Γ (randomized or deterministic):*

1. Γ_{norm} is at least as good as Γ , in the sense that

$$\Gamma_{\text{norm}}^{\epsilon}(x_1, y_1, \dots, x_n) \subseteq \Gamma^{\epsilon}(x_1, y_1, \dots, x_n)$$

for all x_1, y_1, \dots, x_n and ϵ ;

2. $(\Gamma_{\text{norm}})_{\text{norm}} = \Gamma_{\text{norm}}$;
3. Γ is normalized if and only if the set

$$\{\epsilon : y_n \in \Gamma^{\epsilon}(x_1, y_1, \dots, x_n)\}$$

is open in $(0, 1)$ for all x_1, y_1, \dots, x_n ;

4. If Γ is exact (resp. conservative, resp. strongly exact), then Γ_{norm} is exact (resp. conservative, resp. strongly exact).

Proof. All (smoothed) conformal predictors are normalized because all inequalities involving ϵ in (2.18) and (2.20) are strict. Properties 1 and 2 are obvious. Property 3 follows from the following restatement of the definition of Γ_{norm} :

$$y_n \in \Gamma_{\text{norm}}^{\epsilon}(x_1, y_1, \dots, x_n)$$

if and only if

$$\exists \epsilon' > \epsilon : y_n \in \Gamma^{\epsilon'}(x_1, y_1, \dots, x_n).$$

Property 4 follows from

$$\text{err}_n^{\epsilon}(\Gamma_{\text{norm}}) = \inf_{\epsilon' > \epsilon} \text{err}_n^{\epsilon'}(\Gamma). \quad \square$$

The next proposition asserts the equivalence of confidence transducers and normalized confidence predictors (in particular, the equivalence of conformal transducers and conformal predictors). Remember that f' is the confidence predictor associated with a confidence transducer f and Γ' is the confidence transducer associated with a confidence predictor Γ .

Proposition 2.12. *For each confidence transducer f , the confidence predictor f' is normalized and $f'' = f$. For each normalized confidence predictor Γ , $\Gamma'' = \Gamma$. If $\Gamma = f'$ is the normalized confidence predictor associated with a confidence transducer $f = \Gamma'$,*

$$\text{err}_n^\epsilon(\Gamma, (z_1, z_2, \dots)) = \mathbb{I}_{f(z_1, \dots, z_n) \leq \epsilon} \quad (2.51)$$

for any data sequence z_1, z_2, \dots . If a randomized confidence transducer f is exact, f' is strongly exact. If a randomized confidence predictor Γ is strongly exact, Γ' is exact. If f is a (smoothed) conformal transducer, f' is a (smoothed) conformal predictor. If Γ is a (smoothed) conformal predictor, Γ' is a (smoothed) conformal transducer.

Proof. Most of the statements in the proposition are obvious. Equality (2.51) follows from the definition of f' . This equality implies that the confidence transducer Γ' is exact for any strongly exact randomized confidence predictor Γ . Indeed, the p-values p_n output by Γ' have the uniform distribution \mathbf{U} on $[0, 1]$, and it remains to apply the following simple fact: if a sequence p_1, p_2, \dots of random variables distributed as \mathbf{U} is such that, for any sequence $(\epsilon_1, \epsilon_2, \dots) \in (0, 1)^\infty$, the random variables $\mathbb{I}_{p_n \leq \epsilon_n}$, $n = 1, 2, \dots$, are independent, then the random variables p_n themselves are independent (see, e.g., Shiryaev 1996, the theorem in §II.5, p. 179). \square

2.6 Proofs

Proof of Theorem 2.1

The proof will show that no confidence predictor satisfies even the property of being *weakly exact*, where the requirement that $\text{err}_n^\epsilon(\Gamma)$ be independent for different n is dropped and the exchangeability assumption is replaced by the randomness assumption. Moreover, we will see that even for a fixed $n \in \mathbb{N}$ it is impossible to have the probability of $\text{err}_n^\epsilon(\Gamma) = 1$ equal to ϵ for all $\epsilon \in (0, 1)$.

We may assume that the examples z_1, z_2, \dots are generated from a power distribution Q^∞ such that the probability distribution Q on \mathbf{Z} is concentrated on the set $\{(x, y^{(1)}), (x, y^{(2)})\} \subseteq \mathbf{Z}$, for some arbitrarily fixed $x \in \mathbf{X}$ and $y^{(1)}, y^{(2)} \in \mathbf{Y}$ (remember that we assumed $|\mathbf{X}| \geq 1$ and $|\mathbf{Y}| > 1$). Therefore, we assume, without loss of generality, that $\mathbf{Z} = \{0, 1\}$. Fix $n \in \mathbb{N}$ and suppose that $\text{err}_n^\epsilon(\Gamma) = 1$ with probability ϵ for all $\epsilon \in (0, 1)$.

For each k let $f(k)$ be the probability that $\text{err}_n^\epsilon(\Gamma, (z_1, \dots, z_n)) = 1$ (we drop z_{n+1}, z_{n+2}, \dots from our notation since err_n^ϵ does not depend on them)

where $(z_1, \dots, z_n) \in \{0, 1\}^n$ is drawn from the uniform distribution on the set of all binary sequences of length n with k 1s. Since the expected value of $f(k)$ is ϵ w.r. to any binomial distribution on $\{0, 1, \dots, n\}$, the standard completeness result (see, e.g., Lehmann 1986, §4.3) implies that $f(k) = \epsilon$ for all $k = 0, 1, \dots, n$. Therefore, $\epsilon \binom{n}{k}$ is an integer for all k and ϵ , which cannot be true.

Proof of Theorem 2.6

Let $\epsilon \in (0, 1)$, $n \in \mathbb{N}$, Q be a probability distribution on \mathbf{Z} , and $\eta_n^{(\epsilon)}$ and $\xi_n^{(\epsilon)}$ be the random variables from the definition of the conservative validity of Γ corresponding to the exchangeable probability distribution Q^∞ . For each sequence of examples $(z_1, \dots, z_n) \in \mathbf{Z}^n$ let $f(z_1, \dots, z_n)$ be the conditional probability under Q^∞ that $\xi_n^{(\epsilon)} = 1$ given $\eta_i^{(\epsilon)} = \text{err}_i^{(\epsilon)}(\Gamma, (z_1, z_2, \dots))$, $i = 1, \dots, n$. For each bag $B \in \mathbf{Z}^{(n)}$ let $f(B)$ be the arithmetic mean of $f(z_1, \dots, z_n)$ over all $n!$ orderings of B . We know that the expected value of $f(B)$ is ϵ under any Q^n , and this, by the completeness of the statistic that maps data sequences (z_1, \dots, z_n) to bags $\{z_1, \dots, z_n\}$ (see Lehmann 1986, §4.3; since \mathbf{Z} is Borel, it can as well be taken to be \mathbb{R}), implies that $f(B) = \epsilon$ for almost all (under any Q^∞) bags B . Let us only consider such bags. Define $S(B, \epsilon)$ as the bag of elements z of B such that Γ makes an error at significance level ϵ at trial n when fed with the elements of B ordered in such a way that the n th example is z (since Γ is invariant, whether an error is made depends only on which element is last, not on the ordering of the first $n - 1$ elements). It is clear that

$$\epsilon_1 \leq \epsilon_2 \implies S(B, \epsilon_1) \subseteq S(B, \epsilon_2)$$

and

$$\frac{|S(B, \epsilon)|}{n} \leq \epsilon.$$

Therefore, the conformal predictor determined by the conformity measure

$$A_n(B \setminus \{z\}, z) := \inf\{\epsilon : z \in S(B, \epsilon)\}$$

is at least as good as Γ .

Proof of Proposition 2.7

The proof is similar to the proof of the previous proposition but more complicated since it depends on the proof of Proposition 2.4. Let $\epsilon \in (0, 1)$ and, for each $n = 1, 2, \dots$ and each bag $B \in \mathbf{Z}^{(n)}$, let $f(B)$ be the probability that $\text{err}_n^{(\epsilon)}(\Gamma) = 1$ when Γ is supplied with the elements of B in a random order (each order having the same probability $1/n!$) and with random numbers τ_1, \dots, τ_n distributed independently according to \mathbf{U} . Since the expected value of $f(B)$ is ϵ for any power distribution Q^∞ on \mathbf{Z}^∞ generating the examples, the same completeness argument shows that $f(B) = \epsilon$ for almost all bags B . It remains to combine this with the invariance of Γ and the proof of Theorem 8.2 (generalization of Proposition 2.4) in §8.7.

2.7 Bibliographical and historical remarks

Conformal prediction

Conformal predictors were first described by Vovk et al. (June 1999) and Saunders et al. (1999). The independence of errors in the on-line mode was proved in Vovk 2002b.

The idea of a rudimentary conformal predictor based on Vapnik's support vector machine was described by Gammerman et al. (January 1997) and originated at the meeting (mentioned on p. XV) between Gammerman, Vapnik and Vovk in the summer of 1996. Having had worked for a long time on the algorithmic theory of randomness (the paper Vovk and V'yugin 1993 being most relevant), Vovk realized that Chervonenkis's old idea (dating from June 1966, according to Chervonenkis 2004) that a small number of support vectors translates into confident predictions (cf. (10.6)) can be used for making hedged predictions. Let us consider the problem of classification ($|\mathbf{Y}|$ is finite and small). From the point of view of the algorithmic theory of randomness, we can make a confident prediction for the label y_n of the new object x_n given a training set z_1, \dots, z_{n-1} if the algorithmic randomness deficiency is small for only one possible extension $(z_1, \dots, z_{n-1}, (x_n, y))$, $y \in \mathbf{Y}$; we can then output the corresponding y as a confident prediction. (In the case of regression, $\mathbf{Y} = \mathbb{R}$, a confident prediction for x_n is possible if the algorithmic randomness deficiency is small for a narrow range of $y \in \mathbf{Y}$.)

Remark The reader who is not familiar with the algorithmic theory of randomness (which is not used in this book outside the end-of-chapter remarks) can consult Kolmogorov 1983, Martin-Löf 1966, V'yugin 1994, Li and Vitányi 1997. In the literature on algorithmic randomness the word "algorithmic" is often omitted, but we will always keep it, to avoid confusion with several other, unrelated, notions of randomness used in this book. (In particular, there are no obvious connections between algorithmic randomness and the assumption of randomness discussed in the previous chapter.) The algorithmic notion randomness formalizes the intuitive notion of typicalness: an object $\omega \in \Omega$ is regarded as typical of a probability distribution P on Ω (we will also say "under P " or "w.r. to P ") if there is no reason to be surprised when told that ω was drawn randomly from P . Using the notion of a universal Turing machine, it is possible to introduce the notion of *algorithmic randomness deficiency*, formalizing the degree of deviation from typicalness. For further details, see p. 218.

There are two very different approaches to defining algorithmic randomness deficiency: Martin-Löf's (1966) and Levin's (1976, 1984; simplified in Gacs 1980 and Vovk and V'yugin 1993). Kolmogorov's (1968) original definition is a special case of Martin-Löf's, but becomes a special case of Levin's (as simplified in Vovk and V'yugin 1993) if the plain Kolmogorov complexity in it is replaced by prefix complexity. Martin-Löf's definition is more intuitive, being a universal version of the standard statistical notion of p-value, but Levin's definition often leads to more elegant mathematical results.

The paper by Gammerman et al. (1997) was based on Levin's definition, which made it difficult to understand. Conformal predictors, which appeared in Vovk et al. 1999 and Saunders et al. 1999, were the result of replacing Levin's definition of algorithmic randomness by Martin-Löf's definition in Gammerman et al. 1997.

After the notion of conformal predictor crystallized, the connection with the algorithmic theory of randomness started to disappear; in particular, in order to obtain the strongest possible results, we replaced the algorithmic notion of randomness with statistical tests. As we said earlier, in this book we hardly ever mention algorithmic theory of randomness outside the end-of-chapter remarks. This evolution does not look surprising: e.g., we argued in Vovk and Shafer 2003 that the algorithmic notions of randomness and complexity are powerful sources of intuition, but for stating mathematical results in their strongest and most elegant form it is often necessary to “translate” them into a non-algorithmic form.

For the information on the many precursors of conformal prediction, see §10.2; Kei Takeuchi’s definition is especially close to ours.

A version (more sophisticated but less precise, involving arbitrary constants) of Theorem 2.6 was stated and proved in Nouretdinov et al. 2003. An analogous result was stated by Takeuchi for his version of conformal predictors.

Least squares and ridge regression

The least squares procedure was invented independently by Gauss and Legendre and first published by Legendre in 1805 (for details, see, e.g., Plackett 1972, Stigler 1981, 1986a). The term “hat matrix” was introduced by John W. Tukey (see Hoaglin and Welsch 1978). The ridge regression procedure was first described in detail by Arthur E. Hoerl and Robert W. Kennard (1970a, 1970b). The idea came from Hoerl’s (1959) ridge analysis, a method of examining high-dimensional quadratic response surfaces (for details, see Hoerl 1985). The link between ridge analysis and ridge regression is provided by Hoerl’s 1962 paper.

Deleted residuals are also known as PRESS and predicted residuals, and the nonconformity scores (2.36) differ from “internally studentized residuals” only by a factor that does not affect the conformal predictor’s output. We did not consider “externally studentized residuals”; for details and history, see, e.g., Cook and Weisberg 1982 (§2.2.1).

The RRCM was developed by Ilia Nouretdinov and published in Nouretdinov et al. 2001a.

Kernel methods

Kernel methods have their origins in the Hilbert–Schmidt theory of integral equations (see Mercer 1909). The fundamental fact that each symmetric nonnegative definite function has representation (2.42) can be proved by many different methods: see, e.g., Mercer 1909 (that paper, however, proves a slightly different result, “Mercer’s theorem”, about continuous kernels and an integral analogue of condition (2.46)) and Aronszajn 1950 (Aronszajn’s proof is based on Moore’s idea; it is reproduced in Wahba 1990). For recent expositions, see, e.g., Schölkopf and Smola 2002; Cristianini and Shawe-Taylor 2000; Shawe-Taylor and Cristianini 2004.

There are several approaches to the kernel ridge regression; the three main ones appear to be the following:

- the approach adopted in this book: the objects are mapped to an arbitrary (not necessarily functional or separable) Hilbert space and the prediction rule is chosen from among the continuous linear functionals on that space; the main

equation (2.41) can be obtained, for example, using the Lagrange method analogously to Vapnik's (1998) derivation of SVM (see Saunders et al. 1998); the approach based on the equality (2.39) is standard in statistics;

- the approach based on functional Hilbert spaces with bounded evaluation functionals (called *reproducing kernel Hilbert spaces*; the prediction rule is chosen from among the elements of such a space; see, e.g., Wahba 1990);
- the approach based on Gaussian processes: one assumes that the labels y_i are obtained from zero-mean normal random variables with covariances $\text{cov}(y_i, y_j)$ defined in terms of $\mathcal{K}(x_i, x_j)$; (2.41) can then be obtained as the expected value of the x 's label. In geostatistics this approach is known as kriging; for further details, see Cressie 1993 and p. 273.

Formula (2.44), which we used for the fast updating of the inverse matrix in the kernel RRCM, may have been first explicitly given by Banachiewicz (1937a, 1937b); further references and history can be found in Henderson and Searle 1981. There are similar updating formulas (going back to Gauss 1823 and also reviewed in Henderson and Searle 1981) that could be used in the case of RRCM, but the need for speeding up computations is less pressing for RRCM since the matrix to be inverted is always of the constant size $p \times p$.

Classification with conformal predictors

In this chapter we concentrate on the problem of classification, where the label space \mathbf{Y} is finite. We start in §3.1 by giving two more examples of nonconformity measures, this time specifically for the case of classification, and reporting on the empirical performance of one of them. In the next section, §3.2, we state the main result of the chapter: there exists a “universal” smoothed conformal predictor whose asymptotic efficiency¹ is not worse than that of any other asymptotically valid randomized confidence predictor, regardless of the probability distribution Q generating individual examples. In particular, even if for a given probability distribution Q we construct the optimal, or “Bayes”², confidence predictor Γ , our universal predictor will be as efficient as Γ asymptotically, even though the former “knows nothing” about Q . In §3.4 we make the first step towards the proof of the main result looking closely at the Bayes confidence predictor; this will allow us to set the target for the universal predictor. The universal smoothed conformal predictor is constructed in §3.3. As usual, most of the actual proofs will be given in a separate section, §3.5.

The learning protocol of this chapter is the same as in Chap. 2, but we will state it again this time including not only the variables Err_n^ϵ (the total number of errors made up to and including trial n at significance level ϵ) and err_n^ϵ (the binary variable showing whether an error is made at trial n) but also the analogous variables Mult_n^ϵ , mult_n^ϵ , Emp_n^ϵ , emp_n^ϵ for multiple (containing more than one label) and empty (containing no labels) predictions:

$$\text{Err}_0^\epsilon := 0, \text{Mult}_0^\epsilon := 0, \text{Emp}_0^\epsilon := 0 \text{ for all } \epsilon \in (0, 1);$$

FOR $n = 1, 2, \dots$:

Reality outputs $x_n \in \mathbf{X}$;

¹We use the expression “asymptotic efficiency” only informally, to refer to the asymptotic optimality of the predictions made. In this book we never consider the (very interesting) question of how fast optimality is approached.

²We are following standard usage (see Devroye et al. 1996), despite the lack of connection with Bayesian learning (as discussed in, e.g., Chap. 10).

Predictor outputs $\Gamma_n^\epsilon \subseteq \mathbf{Y}$ for all $\epsilon \in (0, 1)$;

Reality outputs $y_n \in \mathbf{Y}$;

$$\text{err}_n^\epsilon := \begin{cases} 1 & \text{if } y_n \notin \Gamma_n^\epsilon \\ 0 & \text{otherwise} \end{cases} \text{ for all } \epsilon \in (0, 1);$$

$\text{Err}_n^\epsilon := \text{Err}_{n-1}^\epsilon + \text{err}_n^\epsilon$ for all $\epsilon \in (0, 1)$;

$$\text{mult}_n^\epsilon := \begin{cases} 1 & \text{if } |\Gamma_n^\epsilon| > 1 \\ 0 & \text{otherwise} \end{cases} \text{ for all } \epsilon \in (0, 1);$$

$\text{Mult}_n^\epsilon := \text{Mult}_{n-1}^\epsilon + \text{mult}_n^\epsilon$ for all $\epsilon \in (0, 1)$;

$$\text{emp}_n^\epsilon := \begin{cases} 1 & \text{if } |\Gamma_n^\epsilon| = 0 \\ 0 & \text{otherwise} \end{cases} \text{ for all } \epsilon \in (0, 1);$$

$\text{Emp}_n^\epsilon := \text{Emp}_{n-1}^\epsilon + \text{emp}_n^\epsilon$ for all $\epsilon \in (0, 1)$

END FOR.

In this chapter, the label space \mathbf{Y} is finite and equipped with the discrete σ -algebra.

3.1 More ways of computing nonconformity scores

First of all we notice that the general scheme discussed at the end of §2.2 is applicable generally, including the case of classification. For classification, it is especially important to allow the case $\hat{\mathbf{Y}} \neq \mathbf{Y}$.

Another general remark is that any procedure of computing nonconformity scores for regression can be used for computing nonconformity scores in binary classification (and there are standard ways to reduce general classification problems to binary ones, as we will see at the end of this section). Indeed, if \mathbf{Y} consists of just two elements, we can encode them by two different real numbers and run the regression procedure for computing nonconformity scores. In particular, we can use the nonconformity scores produced by ridge regression and by nearest neighbors regression, as discussed in the previous chapter, in classification problems.

Nonconformity scores from nearest neighbors

There is, however, a much more direct way of applying the nearest neighbors idea to obtain a nonconformity measure: assuming the objects are vectors in a Euclidean space, the nonconformity scores can be defined, in the spirit of the 1-nearest neighbor algorithm, as

$$A(\{(x_1, y_1), \dots, (x_l, y_l)\}, (x, y)) := \frac{\min_{i=1, \dots, l: y_i=y} d(x, x_i)}{\min_{i=1, \dots, l: y_i \neq y} d(x, x_i)}, \quad (3.1)$$

where d is the Euclidean distance (i.e., an object is considered nonconforming if it is close to an object labeled in a different way and far from any object

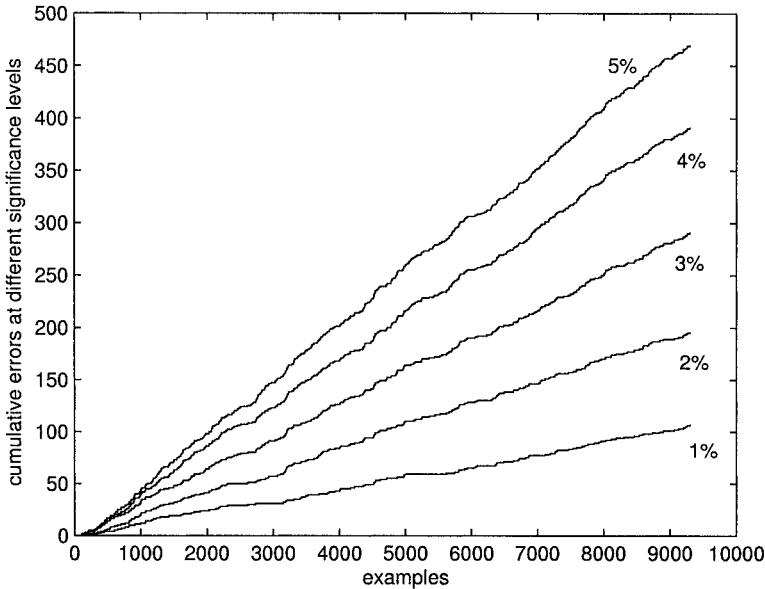


Fig. 3.1. Cumulative errors Err_n^ϵ suffered by the 1-nearest neighbor conformal predictor on the USPS data set (9298 hand-written digits, randomly permuted) plotted against n for the significance levels from $\epsilon = 1\%$ to $\epsilon = 5\%$

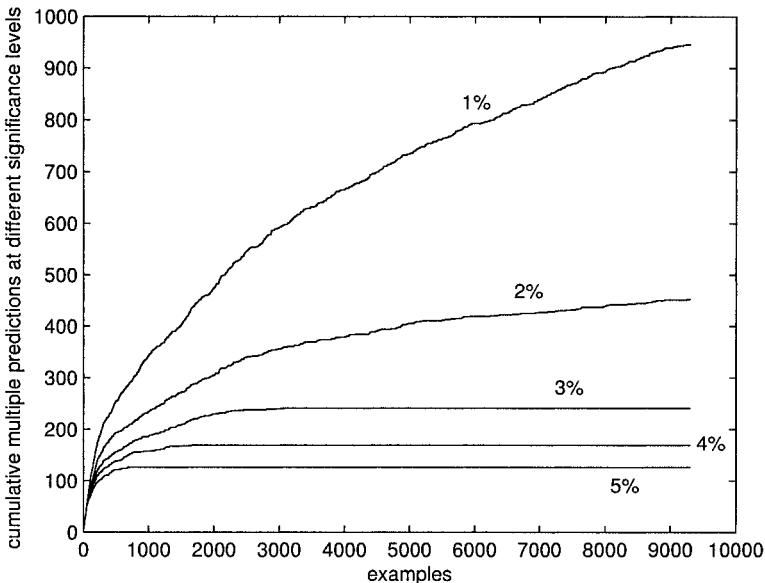


Fig. 3.2. Cumulative number of multiple predictions Mult_n^ϵ output by the 1-nearest neighbor conformal predictor on the USPS data set

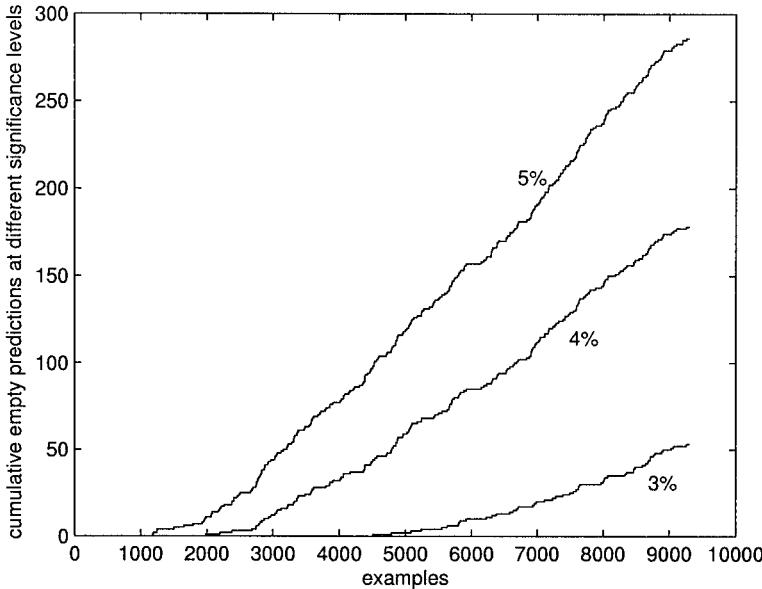


Fig. 3.3. Cumulative number of empty predictions Emp_n^ϵ output by the 1-nearest neighbor conformal predictor on the USPS data set

labeled in the same way). It is possible for (3.1) to be equal to ∞ (if the denominator in (3.1) is zero).

Figures 3.1–3.3 show the on-line performance of the *1-nearest neighbor conformal predictor* (determined by (3.1)) on the USPS data set (the original 9298 hand-written digits, as described in Appendix B, but randomly permuted) for the confidence levels 95–99%. Figure 3.1 again confirms empirically the validity of conformal predictors; the cumulative numbers of errors at $\epsilon = 1\%$ and $\epsilon = 5\%$ were already given in Chap. 1 (Fig. 1.5 on p. 10). Figures 3.2 and 3.3 show that for a vast majority of examples the prediction set contains precisely one label at the considered significance levels. Figure 3.4 illustrates a feature of Figs. 3.2 and 3.3 that is not very noticeable since it requires examination of both figures simultaneously: at a fixed significance level, empty predictions appear only after multiple predictions disappear. (This figure cannot be directly compared to the error rate of 2.5% for humans reported in Vapnik 1998, since our experiment has been carried out on the randomly permuted data set, whereas the test part of the USPS data set is known to be especially hard.)

Nonconformity scores from support vector machines

Support vector machines were proposed by Vapnik (1998, Part II); the standard abbreviation of “support vector machine” is SVM. We concentrate on the problem of binary classification, assuming that the set \mathbf{Y} of possible labels

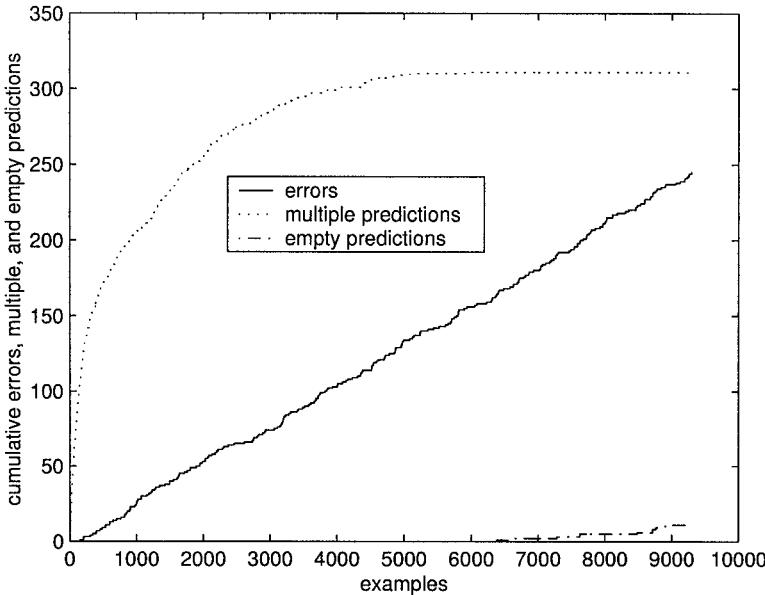


Fig. 3.4. On-line performance of the 1-nearest neighbor conformal predictor on the USPS data set for the significance level 2.5%. The solid line shows the cumulative number of errors, dotted the cumulative number of multiple predictions, and dashdot the cumulative number of empty predictions

is $\{-1, 1\}$. For a given bag

$$\{z_1, \dots, z_n\} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

and its element z_i , each SVM, to be defined shortly, provides a very natural definition of the nonconformity score

$$\alpha_i = A_n(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}, z_i). \quad (3.2)$$

Suppose the objects are vectors in a dot product space \mathbf{H} and consider the quadratic optimization problem

$$\frac{1}{2}(w \cdot w) + C \left(\sum_{i=1}^n \xi_i \right) \rightarrow \min, \quad (3.3)$$

where C is an *a priori* fixed positive constant and the variables $w \in \mathbf{H}$, $\xi = (\xi_1, \dots, \xi_n)' \in \mathbb{R}^n$, $b \in \mathbb{R}$ (the last variable not entering (3.3)) are subject to the constraints

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, \dots, n, \quad (3.4)$$

$$\xi_i \geq 0, \quad i = 1, \dots, n. \quad (3.5)$$

If this optimization problem has a solution, it is unique and is also denoted $w, (\xi_1, \dots, \xi_n)', b$. The hyperplane $w \cdot x + b = 0$, called the *optimal separating hyperplane*, is the boundary of the prediction rule produced by the corresponding SVM: the prediction \hat{y} for a new object x is 1 if $w \cdot x + b > 0$ and -1 if $w \cdot x + b < 0$ (with an arbitrary convention if $w \cdot x + b = 0$).

The next step in the development of SVM is to consider an arbitrary object space \mathbf{X} (not necessarily a linear space) and apply a transformation $F : \mathbf{X} \rightarrow \mathbf{H}$ mapping the objects x_i into the “feature vectors” $F(x_i) \in \mathbf{H}$, where \mathbf{H} is a dot product space. This replaces x_i by $F(x_i)$ in the optimization problem (3.3)–(3.5); as before, w ranges over \mathbf{H} , but the latter is now different from the object space. After that the Lagrange method is applied to the modified problem; to each inequality in (3.4) corresponds a Lagrange multiplier α_i . The optimal values of α_i , obtained by solving the *dual problem*

$$\begin{aligned} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j \alpha_i \alpha_j \mathcal{K}(x_i, x_j) &\rightarrow \max, \\ \sum_{i=1}^n y_i \alpha_i &= 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n, \end{aligned} \tag{3.6}$$

where $\mathcal{K}(x_i, x_j) := F(x_i) \cdot F(x_j)$, can be interpreted as follows:

- the examples z_i with $\alpha_i = 0$ are typical;
- the examples with $\alpha_i = C$ are the most extreme (under the given choice of C) outliers;
- the examples with $0 < \alpha_i < C$ are intermediate, with a possible interpretation of α_i as a measure of nonconformity of the corresponding example.

This makes the solutions to the dual problem ideal for use as nonconformity scores (3.2).

Remark Actually, it is quite possible that the Lagrange multipliers computed by a given computer implementation of SVM will not provide a bona fide nonconformity measure, with α_i in (3.2) depending on the order in which the examples $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ are presented. The order of the examples may be especially important in so-called “chunking”, a standard feature of SVM implementations. To ensure the invariance of α_i w.r. to permutations of $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ (and so the validity of the resulting conformal predictor), the examples z_1, \dots, z_n can be sorted in some way (e.g., in the lexicographic order of their ASCII representations) obtaining $z_{\pi(1)}, \dots, z_{\pi(n)}$, where π is a permutation of the set $\{1, \dots, n\}$. The Lagrange multipliers computed from $z_{\pi(1)}, \dots, z_{\pi(n)}$ should then be permuted using π^{-1} to obtain $\alpha_1, \dots, \alpha_n$.

Reducing classification problems to the binary case

The original SVM method can only deal with binary classification problems, but we will now see that there are ways to use it for solving *multilabel* classification problems (i.e., those with $|\mathbf{Y}| > 2$).

There are two standard ways to reduce multilabel classification problems to the binary case: the “one-against-the-rest” procedure and the “one-against-one” procedure. Suppose we have a reasonable nonconformity measure A for binary classification but are confronted with a multilabel classification problem. For concreteness we will assume that the label space in the binary classification problem is $\{0, 1\}$; if it is $\{a, b\}$ (e.g., $a = -1$ and $b = 1$ in the case of SVM), the reduction will be achieved by a further scaling, $y \mapsto a + (b - a)y$.

The *one-against-the-rest procedure* gives the nonconformity measure

$$\begin{aligned} A^{(1)}(\{(x_1, y_1), \dots, (x_l, y_l)\}, (x, y)) \\ := \lambda A(\{(x_1, \mathbb{I}_{y_1=y}), \dots, (x_l, \mathbb{I}_{y_l=y})\}, (x, 1)) \\ + \frac{1-\lambda}{|\mathbf{Y}|-1} \sum_{y' \neq y} A(\{(x_1, \mathbb{I}_{y_1=y'}), \dots, (x_l, \mathbb{I}_{y_l=y'})\}, (x, 0)), \end{aligned} \quad (3.7)$$

where $\lambda \in [0, 1]$ is a constant (parameter of the procedure) and \mathbb{I} is the indicator function (i.e., $\mathbb{I}_E = 1$ if E holds and $\mathbb{I}_E = 0$ if not). Intuitively, we consider $|\mathbf{Y}|$ auxiliary binary classification problems and compute the nonconformity score of an example (x, y) as the weighted average of the scores this example receives in the auxiliary problems.

The *one-against-one procedure* gives the nonconformity measure

$$A^{(2)}(\{(x_1, y_1), \dots, (x_l, y_l)\}, (x, y)) := \frac{1}{|\mathbf{Y}|-1} \sum_{y' \neq y} A(B_{y,y'}, (x, 1)), \quad (3.8)$$

where $B_{y,y'}$ is the bag obtained from $\{(x_1, y_1), \dots, (x_l, y_l)\}$ as follows: remove all (x_i, y_i) with $y_i \notin \{y, y'\}$; replace each (x_i, y) by $(x_i, 1)$; replace each (x_i, y') by $(x_i, 0)$. We now have $|\mathbf{Y}| - 1$ auxiliary binary classification problems.

The numbers $|\mathbf{Y}|$ and $|\mathbf{Y}| - 1$ of auxiliary binary classification problems given above refer to computing only one nonconformity score. When using nonconformity measures for conformal prediction, we have to compute all n nonconformity scores (2.19) (p. 26) for all $y \in \mathbf{Y}$. With the one-against-the-rest procedure, we have to consider $2|\mathbf{Y}|$ auxiliary binary classification problems altogether, whereas with the one-against-one procedure $|\mathbf{Y}|(|\mathbf{Y}| - 1)$ auxiliary binary classification problems are required. When $|\mathbf{Y}| = 3$, the numbers of auxiliary problems coincide, $2|\mathbf{Y}| = |\mathbf{Y}|(|\mathbf{Y}| - 1)$, but for $|\mathbf{Y}| > 3$ the one-against-the-rest procedure requires fewer auxiliary problems, $2|\mathbf{Y}| < |\mathbf{Y}|(|\mathbf{Y}| - 1)$.

3.2 Universal predictor

We first describe the main idea of a universal confidence predictor. Let us fix an exchangeable probability distribution P on \mathbf{Z}^∞ generating the examples z_1, z_2, \dots , and let us fix a significance level ϵ . Remember that \mathbf{Y} is finite and $|\mathbf{Y}| > 1$.

Slightly elaborating on the notion introduced in Chap. 2, we say that a confidence predictor is *asymptotically conservative for P* and ϵ if the long-run frequency of errors does not exceed ϵ almost surely w.r. to P ; we know that each conformal predictor satisfies this property. For asymptotically conservative predictors we take the number Mult_n of multiple predictions as the principal measure of predictive efficiency. The main result of this chapter is the construction of a confidence predictor (a smoothed conformal predictor) which, for any (unknown) P and any ϵ : (a) makes errors independently and with probability ϵ at every trial (in particular, is asymptotically conservative for P and ϵ); (b) makes in the long run no more multiple predictions than any other randomized confidence predictor that is asymptotically conservative for P and ϵ ; (c) processes example n in time $O(\log n)$.

There is a slight complication for item (b), dealing with predictive efficiency: we also have to deal carefully with empty predictions. The full picture is that our universal predictor, for any significance level ϵ and without knowing the true distribution P generating the examples:

- produces, asymptotically, no more multiple predictions than any other randomized confidence predictor that is asymptotically conservative for P and ϵ ;
- produces, asymptotically, at least as many empty predictions as any other randomized confidence predictor that is asymptotically conservative for P and ϵ and whose percentage of multiple predictions is optimal (in the sense of the previous item).

The importance of the first item is obvious: we want to minimize the number of multiple predictions. This criterion ceases to work, however, when the number of multiple predictions stabilizes, as in the case of significance levels 3%–5% in Fig. 3.2. In such cases the number of empty predictions becomes important: empty predictions (automatically leading to an error) provide a warning that the object is untypical (looks very different from the previous objects), and one would like to be warned as often as possible, taking into account that the frequency of errors (including empty predictions) is guaranteed not to exceed ϵ in the long run.

We now start the formal exposition, only considering randomized confidence predictors. We will often use the notation mult_n^ϵ , emp_n^ϵ , etc., in the case where Predictor and Reality are using given randomized strategies, as was already done in the previous chapter for err_n^ϵ and Err_n^ϵ ; for example, $\text{mult}_n^\epsilon(\Gamma, P)$ is the random variable equal to 1 if Predictor makes a multiple prediction at trial n and 0 otherwise. It is always assumed that the random numbers τ_n used by Γ and the random examples z_n chosen by Reality are independent.

We say that a randomized confidence predictor Γ is *asymptotically conservative for a probability distribution P on \mathbf{Z}^∞ and a significance level $\epsilon \in (0, 1)$* if

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon(\Gamma, P)}{n} \leq \epsilon \quad \text{a.s.}$$

We say that Γ is *asymptotically optimal* for P and ϵ if, for any randomized confidence predictor Γ^\dagger which is asymptotically conservative for P and ϵ ,

$$\limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, P)}{n} \leq \liminf_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma^\dagger, P)}{n} \quad \text{a.s.} \quad (3.9)$$

(It is natural to assume in this and other similar definitions that the random numbers used by Γ and Γ^\dagger are independent, but this assumption is not needed for our mathematical results and we do not make it.) Of course, the definition of asymptotic optimality is natural only for asymptotically conservative Γ .

A randomized confidence predictor Γ is a *universal predictor* if:

- it is asymptotically conservative for any exchangeable P and ϵ ;
- it is asymptotically optimal for any exchangeable P and ϵ ;
- for any exchangeable P , any ϵ , and any randomized confidence predictor Γ^\dagger which is asymptotically conservative and optimal for P and ϵ ,

$$\liminf_{n \rightarrow \infty} \frac{\text{Emp}_n^\epsilon(\Gamma, P)}{n} \geq \limsup_{n \rightarrow \infty} \frac{\text{Emp}_n^\epsilon(\Gamma^\dagger, P)}{n} \quad \text{a.s.}$$

Now we can state the main result of this chapter.

Theorem 3.1. *Suppose the object space \mathbf{X} is Borel. There exists a universal predictor.*

In the next section we construct a universal predictor (processing example n in time $O(\log n)$).

3.3 Construction of a universal predictor

Preliminaries

If τ is a number in $[0, 1]$, we split it into two numbers $\tau', \tau'' \in [0, 1]$ as follows: if the binary expansion of τ is $0.a_1a_2\dots$ (redefine the binary expansion of 1 to be $0.11\dots$), set $\tau' := 0.a_1a_3a_5\dots$ and $\tau'' := 0.a_2a_4a_6\dots$. If τ is distributed uniformly in $[0, 1]$, then both τ' and τ'' are, and they are independent of each other.

In this chapter we will especially often apply our procedures (such as nonconformity measures and prediction rules) not to the original objects $x \in \mathbf{X}$ but to the extended objects $(x, \sigma) \in \tilde{\mathbf{X}} := \mathbf{X} \times [0, 1]$, where x is complemented by a random number σ (to be extracted from one of the τ_n). Along with examples (x, y) we will thus also consider *extended examples* $(x, \sigma, y) \in \tilde{\mathbf{Z}} := \mathbf{X} \times [0, 1] \times \mathbf{Y}$.

Let us set $\mathbf{X} := [0, 1]$; we can do this without loss of generality since \mathbf{X} is Borel. This makes the extended object space $\tilde{\mathbf{X}} = [0, 1]^2$ a linearly ordered

set with the *lexicographic order*: $(x_1, \sigma_1) < (x_2, \sigma_2)$ means that either $x_1 = x_2$ and $\sigma_1 < \sigma_2$ or $x_1 < x_2$. We say that (x_1, σ_1) is *nearer* to (x_3, σ_3) than (x_2, σ_2) is if

$$|x_1 - x_3, \sigma_1 - \sigma_3| < |x_2 - x_3, \sigma_2 - \sigma_3| , \quad (3.10)$$

where

$$|x, \sigma| := \begin{cases} (x, \sigma) & \text{if } (x, \sigma) \geq (0, 0) \\ (-x, -\sigma) & \text{otherwise.} \end{cases}$$

If (x_1, σ_1) and (x_2, σ_2) are extended objects, we will sometimes refer to $|x_1 - x_2, \sigma_1 - \sigma_2|$ as the *distance* between (x_1, σ_1) and (x_2, σ_2) , even though this distance is a two-dimensional object (what is important is that the distances are linearly ordered according to (3.10)).

Our construction will be based on the nearest neighbors algorithm, which is known to be strongly universally consistent in the traditional theory of pattern recognition (see, e.g., Devroye et al. 1996, Chap. 11); the random components σ are needed for tie-breaking. It is still theoretically possible for the expression “the k th nearest neighbor” not to have a precise meaning: two extended objects in a training set can be at the same distance from a given extended object. This case, which happens with probability zero, will be always treated separately.

Conformal prediction in the current context

The smoothed conformal predictors we are going to construct will work on extended examples; otherwise it will be our standard notion. (It might have been better to call them “doubly smoothed conformal predictors”, but we will not make such fine distinctions.) Therefore, a nonconformity measure is a mapping $A : \tilde{\mathbf{Z}}^{(*)} \times \tilde{\mathbf{Z}} \rightarrow \mathbb{R}$. The smoothed conformal predictor determined by the nonconformity measure A is the following randomized confidence predictor

$$\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) : \quad (3.11)$$

at each trial n and for each $y \in \mathbf{Y}$, define

$$\begin{aligned} \alpha_i &:= A \left(\lceil (x_1, \tau'_1, y_1), \dots, (x_{i-1}, \tau'_{i-1}, y_{i-1}), \right. \\ &\quad \left. (x_{i+1}, \tau'_{i+1}, y_{i+1}), \dots, (x_{n-1}, \tau'_{n-1}, y_{n-1}), (x_n, \tau'_n, y) \rceil, (x_i, \tau'_i, y_i) \right), \\ &\quad i = 1, \dots, n-1 , \end{aligned}$$

and

$$\alpha_n := A \left(\lceil (x_1, \tau'_1, y_1), \dots, (x_{n-1}, \tau'_{n-1}, y_{n-1}) \rceil, (x_n, \tau'_n, y) \right) ;$$

include y in (3.11) if and only if

$$\frac{|\{i = 1, \dots, n : \alpha_i > \alpha_n\}| + \tau''_n |\{i = 1, \dots, n : \alpha_i = \alpha_n\}|}{n} > \epsilon . \quad (3.12)$$

We already know that every (smoothed) conformal predictor is asymptotically conservative for every P and ϵ .

Universal predictor

Fix a strictly increasing sequence of integers K_n , $n = 1, 2, \dots$, such that

$$K_n \rightarrow \infty, K_n = o\left(\sqrt{n/\ln n}\right) \quad (3.13)$$

as $n \rightarrow \infty$. Let $B = \{w_1, \dots, w_n\}$ be a bag of extended examples $w_i = (x_i, \sigma_i, y_i)$. The nearest neighbors approximation $\hat{Q}_B(y | x, \sigma)$ to the true (but unknown) conditional probability that an object x 's label is y is defined as

$$\hat{Q}_B(y | x, \sigma) := N_B(x, \sigma, y)/K_n, \quad (3.14)$$

where $n := l + 1$ and $N_B(x, \sigma, y)$ is the number of $i = 1, \dots, l$ such that $y_i = y$ and (x_i, σ_i) is one of the K_n nearest neighbors of (x, σ) in the sequence $((x_1, \sigma_1), \dots, (x_l, \sigma_l))$. If $K_n \geq n$ or $K_n \leq 0$, this definition does not work, so set, e.g., $\hat{Q}_B(y | x, \sigma) := 1/|\mathbf{Y}|$ for all y and (x, σ) (this particular convention is not essential since, by (3.13), $0 < K_n < n$ from some n on). It is also possible that the phrase “ K_n nearest neighbors of (x, σ) ” is not defined because of distance ties; in this case we again set $\hat{Q}_B(y | x, \sigma) := 1/|\mathbf{Y}|$ for all y .

Define the “empirical predictability function” \hat{f}_B by

$$\hat{f}_B(x, \sigma) := \max_{y \in \mathbf{Y}} \hat{Q}_B(y | x, \sigma).$$

For all B and (x, σ) fix some

$$\hat{y}_B(x, \sigma) \in \arg \max_y \hat{Q}_B(y | x, \sigma)$$

(e.g., take the first element of $\arg \max_y \hat{Q}_B(y | x, \sigma)$ in a fixed ordering of \mathbf{Y}) and define the *nearest neighbors nonconformity measure* by

$$A(B, (x, \sigma, y)) := \begin{cases} -\hat{f}_B(x, \sigma) & \text{if } y = \hat{y}_B(x, \sigma) \\ \hat{f}_B(x, \sigma) & \text{otherwise,} \end{cases} \quad (3.15)$$

B ranging over the bags of extended examples. The *nearest neighbors smoothed conformal predictor* is defined to be the smoothed conformal predictor determined by the nearest neighbors nonconformity measure. The nearest neighbors smoothed conformal predictor will later be shown to be universal.

Proposition 3.2. *Let $\{\epsilon_1, \dots, \epsilon_K\} \subseteq (0, 1)$ be a finite set. If $\mathbf{X} = [0, 1]$ and $K_n \rightarrow \infty$ sufficiently slowly, the nearest neighbors smoothed conformal predictor can be implemented for significance levels $\epsilon = \epsilon_1, \dots, \epsilon_K$ so that computations at trial n are performed in time $O(\log n)$.*

Proposition 3.2 assumes a computational model that allows operations (such as comparison) with real numbers. If \mathbf{X} is an arbitrary Borel space, for this proposition to be applicable \mathbf{X} should be embedded in $[0, 1]$ first; e.g., if $\mathbf{X} \subseteq [0, 1]^n$, an $x = (x_1, \dots, x_n) \in \mathbf{X}$ can be represented as

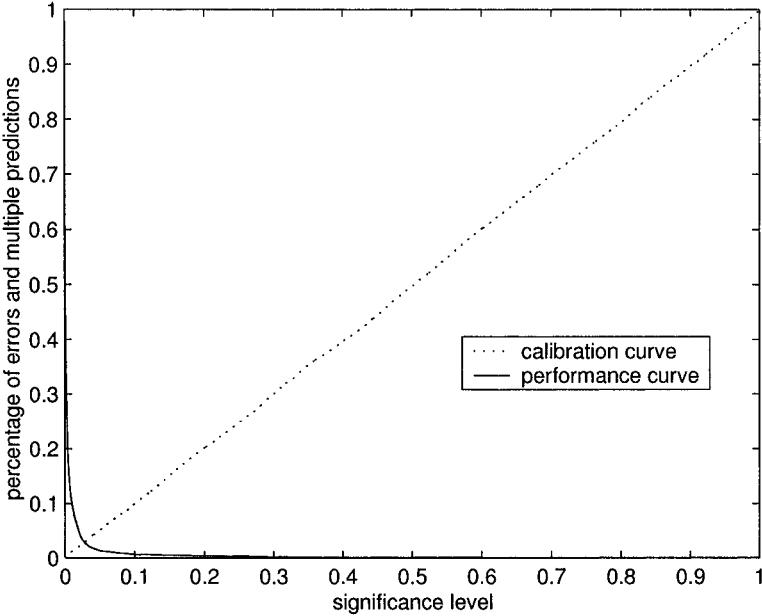


Fig. 3.5. The empirical calibration and performance curves for the 1-nearest neighbor conformal predictor on the USPS data set (randomly permuted)

$$0.x_{i,1}x_{i,2}\dots x_{n,1}x_{1,2}x_{2,2}\dots x_{n,2}\dots \in [0, 1],$$

where $0.x_{i,1}x_{i,2}\dots$ is the binary expansion of x_i . We use the expression “can be implemented” in a wide sense, only requiring that the implementation should give the correct results almost surely.

3.4 Fine details of confidence prediction

In this section we make first steps towards the proof of Theorem 3.1. By de Finetti’s theorem (see §A.5), each exchangeable distribution on \mathbf{Z}^∞ (which is a Borel space as long as \mathbf{Z} is Borel: see, e.g. Schervish 1995, Lemma B.41) is a mixture of power distributions. Therefore, without loss of generality we assume that $P = Q^\infty$ for a probability distribution Q on \mathbf{Z} .

To provide the reader with extra intuition about confidence prediction in the case of classification, we first briefly discuss further empirical results for the 1-nearest neighbor conformal predictor and the USPS data set. Recall that Figs. 3.1–3.3 show the cumulative number of errors, the cumulative number of multiple predictions, and that of empty predictions. Figure 3.5 gives the *empirical calibration curve*

$$\epsilon \mapsto \frac{\text{Err}_N^\epsilon(\Gamma, \text{USPS})}{N} \tag{3.16}$$

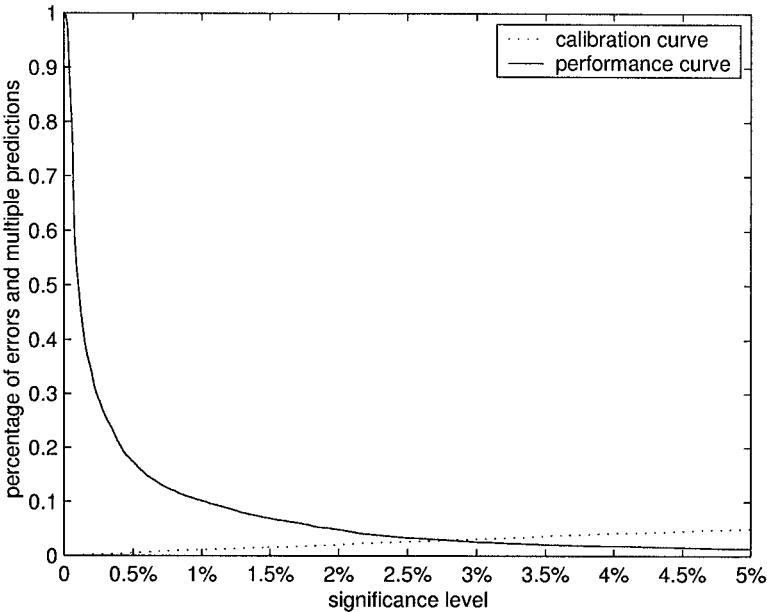


Fig. 3.6. The left edge of the previous figure stretched horizontally

and the *empirical performance curve*

$$\epsilon \mapsto \frac{\text{Mult}_N^\epsilon(\Gamma, \text{USPS})}{N}$$

for this confidence predictor; we use the strategies followed by Reality (the USPS data set, randomly permuted) and Predictor as arguments for Err and Mult. Remember that the size of the USPS data set is $N = 9298$.

We denote by $Q_{\mathbf{X}}$ the marginal distribution of Q on \mathbf{X} (i.e., $Q_{\mathbf{X}}(E) := Q(E \times \mathbf{Y})$) and by $Q_{\mathbf{Y}|\mathbf{X}}(y|x)$ the conditional probability that, for a random example (X, Y) drawn from Q , $Y = y$ provided $X = x$ (we fix arbitrarily a regular version of this conditional probability; the existence of regular conditional probability is obvious in our case of finite \mathbf{Y} and also follows from general results: see §A.3). We will often omit lower indices \mathbf{X} and $\mathbf{Y}|\mathbf{X}$.

The *predictability* of an object $x \in \mathbf{X}$ is

$$f(x) := \max_{y \in \mathbf{Y}} Q(y|x)$$

and the *predictability distribution function* is the increasing³ function $F : [0, 1] \rightarrow [0, 1]$ defined by

³In this book “increasing” and “decreasing” are used in Bourbaki’s weak sense: e.g., $F(\beta)$ is called increasing if $F(\beta_1) \leq F(\beta_2)$ whenever $\beta_1 \leq \beta_2$.

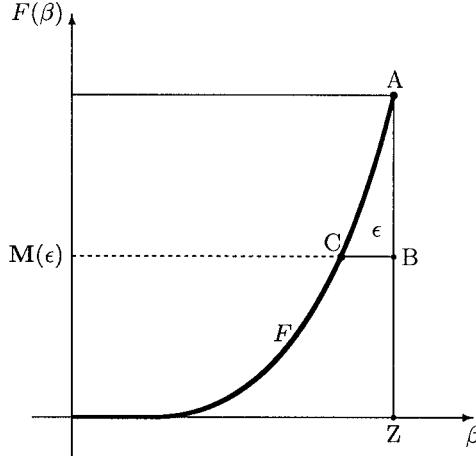


Fig. 3.7. The predictability distribution function F and how it determines the multiplicity curve $M(\epsilon)$. The function F is increasing, continuous on the right, and $F(1/|\mathbf{Y}|^-) = 0$. For a possibly more realistic example of a predictability distribution function, see Fig. 3.12

$$F(\beta) := Q\{x : f(x) \leq \beta\}$$

(essentially, it is the distribution function of the image Qf^{-1} of the probability distribution Q under the mapping f). An example of such a function F is given in Fig. 3.7; the graph of F is the thick line, and the unit box is also shown. (The intuition behind some constructions in this chapter will become clearer if the case of finite \mathbf{X} with equiprobable objects is considered first; see Fig. 3.8.)

The *multiplicity curve* $M = M_Q$ of Q is defined by the equality

$$M(\epsilon) = \inf \left\{ B \in [0, 1] : \int_0^1 (F(\beta) - B)^+ d\beta \leq \epsilon \right\},$$

where t^+ stands for $\max(t, 0)$; the function M is also of the type $[0, 1] \rightarrow [0, 1]$. (Why the terminology introduced here and below is natural will become clear from Propositions 3.3 and 3.5.) Geometrically, M is defined from the graph of F as follows (see Fig. 3.7): move the point B from A to Z until the area of the curvilinear triangle ABC becomes ϵ (assuming this area does become ϵ eventually, i.e., ϵ is not too large); the ordinate of B is then $M(\epsilon)$. The intuition in the case of finite \mathbf{X} (see Fig. 3.8) is that $1 - M(\epsilon)$ is the maximum fraction of objects that are “easily predictable” in the sense that their cumulative lack of predictability does not exceed ϵ (where the lack of predictability $1 - f(x)$ of each object is taken with the weight $1/|\mathbf{X}|$).

The *emptiness curve* $E = E_Q$ of Q is defined by

$$E(\epsilon) = \sup \left\{ B \in [0, 1] : B + \int_0^1 (F(\beta) - B)^+ d\beta \leq \epsilon \right\},$$

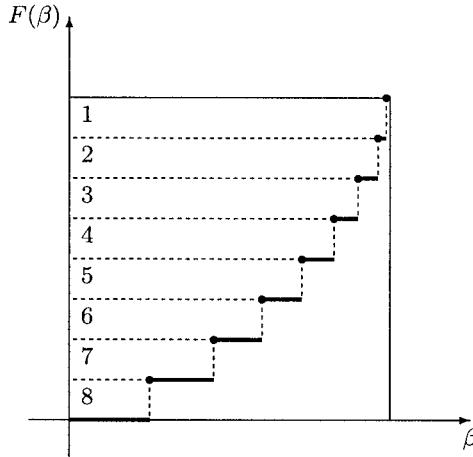


Fig. 3.8. The predictability distribution function (thick line) in the case where the object space \mathbf{X} is finite and all objects $x \in \mathbf{X}$ have the same probability. The objects are numbered, from 1 to 8, in the order of decreasing predictability (equal to the length of the corresponding rectangle)

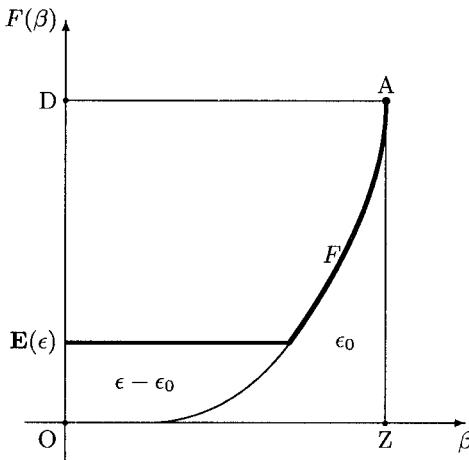


Fig. 3.9. The predictability distribution function F and how it determines the emptiness curve $E(\epsilon)$

with $\sup \emptyset$ interpreted as 0. Similarly to the case of $M(\epsilon)$, $E(\epsilon)$ is defined as the value such that the area of the part of the box AZOD below the thick line in Fig. 3.9 is ϵ ($E(\epsilon) = 0$ if such a value does not exist).

Define the *critical significance level* ϵ_0 as

$$\epsilon_0 := \int_0^1 F(\beta) d\beta \quad (3.17)$$

(the area under the thick curve in Fig. 3.7; we will later see that this coincides with what is sometimes called the *Bayes error* – see, e.g., Devroye et al. 1996, §2.1). It is clear that

$$\begin{aligned} \epsilon \leq \epsilon_0 &\implies \int_0^1 (F(\beta) - \mathbf{M}(\epsilon))^+ d\beta = \epsilon \text{ \& } \mathbf{E}(\epsilon) = 0 \\ \epsilon \geq \epsilon_0 &\implies \mathbf{M}(\epsilon) = 0 \text{ \& } \mathbf{E}(\epsilon) + \int_0^1 (F(\beta) - \mathbf{E}(\epsilon))^+ d\beta = \epsilon. \end{aligned}$$

So far we have defined some characteristics of the distribution Q itself; now we will give definitions pertaining to individual confidence predictors. The most natural class of confidence predictors consists of what we called in Chap. 2 *invariant confidence predictors*: those confidence predictors Γ for which $\Gamma^\epsilon(z_1, \dots, z_l, x)$ does not depend on the order of z_1, \dots, z_l . This includes the definition of randomized confidence predictors as a special case (where z_i range over $\mathbf{X} \times [0, 1] \times \mathbf{Y}$ instead of \mathbf{Z} and x ranges over $\mathbf{X} \times [0, 1]$ instead of \mathbf{X}).

The *calibration curve* of a randomized confidence predictor Γ under Q is the following function of the type $[0, 1] \rightarrow [0, 1]$:

$$\mathbf{C}_{\Gamma, Q}(\epsilon) := \inf \left\{ \beta : \mathbb{P} \left\{ \limsup_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon(\Gamma, Q^\infty)}{n} \leq \beta \right\} = 1 \right\} \quad (3.18)$$

($\mathbb{P}(E)$ stands for the probability of event E). By the Hewitt–Savage zero-one law (see, e.g., Shiryaev 1996, Theorem IV.1.3) in the case of invariant predictors this definition will not change if “= 1” is replaced by “> 0” in (3.18). The *performance curve* of Γ under Q is defined by

$$\mathbf{P}_{\Gamma, Q}(\epsilon) := \inf \left\{ \beta : \mathbb{P} \left\{ \limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \leq \beta \right\} = 1 \right\}; \quad (3.19)$$

this is again a function of the type $[0, 1] \rightarrow [0, 1]$. The Hewitt–Savage zero-one law again implies that for invariant Γ this will not change if “= 1” is replaced by “> 0”.

Notice that a randomized confidence predictor Γ is asymptotically conservative for Q^∞ and any $\epsilon \in (0, 1)$ if its calibration curve $\mathbf{C}_{\Gamma, Q}$ is below the diagonal: $\mathbf{C}_{\Gamma, Q}(\epsilon) \leq \epsilon$ for any significance level ϵ . The next proposition shows that it is asymptotically optimal for Q^∞ and any $\epsilon \in (0, 1)$ if its performance curve coincides with the multiplicity curve: $\mathbf{P}_{\Gamma, Q}(\epsilon) = \mathbf{M}_Q(\epsilon)$ for all ϵ (and we will later see that Γ is asymptotically optimal for Q^∞ and any $\epsilon \in (0, 1)$ only if this condition holds). We will often omit the lower indices in (3.18) and (3.19).

Proposition 3.3. *Let Q be a probability distribution on \mathbf{Z} with the multiplicity curve \mathbf{M} and let $\epsilon \in (0, 1)$. If a randomized confidence predictor Γ is asymptotically conservative for Q^∞ and ϵ , its performance curve $\mathbf{P}_{\Gamma, Q}$ is above \mathbf{M} at ϵ : $\mathbf{P}_{\Gamma, Q}(\epsilon) \geq \mathbf{M}(\epsilon)$. Moreover,*

$$\liminf_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \geq \mathbf{M}(\epsilon) \quad a.s. \quad (3.20)$$

Of course, this proposition will continue to hold if the word “randomized” is omitted. The “a.s.” in (3.20) refers to the probability distribution $(Q \times \mathbf{U})^\infty$ generating the sequence $z_1, \tau_1, z_2, \tau_2, \dots$, with \mathbf{U} standing for the uniform distribution on $[0, 1]$.

Since we are also interested in the number of empty predictions made, we complement Proposition 3.3 with

Proposition 3.4. *Let Q be a probability distribution on \mathbf{Z} with multiplicity curve \mathbf{M} and emptiness curve \mathbf{E} and let $\epsilon \in (0, 1)$ be a significance level. If a randomized confidence predictor Γ is asymptotically conservative for Q and ϵ and satisfies*

$$\limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \leq \mathbf{M}(\epsilon) \quad a.s. , \quad (3.21)$$

then

$$\limsup_{n \rightarrow \infty} \frac{\text{Emp}_n^\epsilon(\Gamma, Q^\infty)}{n} \leq \mathbf{E}(\epsilon) \quad a.s.$$

Theorem 3.1 immediately follows from Propositions 3.3, 3.4 and the following proposition.

Proposition 3.5. *Suppose \mathbf{X} is a Borel space. For any $Q \in \mathbf{P}(\mathbf{Z})$ and any significance level ϵ , the nearest neighbors smoothed conformal predictor Γ constructed in §3.3 satisfies*

$$\limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \leq \mathbf{M}_Q(\epsilon) \quad a.s. \quad (3.22)$$

and

$$\liminf_{n \rightarrow \infty} \frac{\text{Emp}_n^\epsilon(\Gamma, Q^\infty)}{n} \geq \mathbf{E}_Q(\epsilon) \quad a.s. \quad (3.23)$$

Bayes confidence predictor

Let us now assume, for simplicity, that the distribution Q is *regular*, in the sense that the predictability distribution function F is continuous.

In this chapter we prove that one can construct an asymptotically optimal smoothed conformal predictor. If, however, we know for sure that Q is the true distribution on \mathbf{Z} , it is very easy to construct an asymptotically conservative and optimal confidence predictor. Fix a *choice function* $\hat{y} : \mathbf{X} \rightarrow \mathbf{Y}$ such that

$$\forall x \in \mathbf{X} : f(x) = Q(\hat{y}(x) | x) \quad (3.24)$$

(to put it differently, $\hat{y}(x) \in \arg \max_y Q(y | x)$). Define the *Q-Bayes confidence predictor* Γ by

$$\Gamma^\epsilon(z_1, \dots, z_l, x) := \begin{cases} \{\hat{y}(x)\} & \text{if } F(f(x)) \geq \max(\mathbf{M}(\epsilon), \mathbf{E}(\epsilon)) \\ \mathbf{Y} & \text{if } F(f(x)) < \mathbf{M}(\epsilon) \\ \emptyset & \text{if } F(f(x)) < \mathbf{E}(\epsilon) \end{cases}$$

for all significance levels ϵ and data sequences $(z_1, \dots, z_l, x) \in \mathbf{Z}^l \times \mathbf{X}$, $l = 0, 1, \dots$. It can be shown that the *Q-Bayes confidence predictor* is asymptotically conservative and optimal for Q^∞ and any $\epsilon \in (0, 1)$; in addition, it satisfies (3.23) for all $\epsilon \in (0, 1)$ (it also satisfies (3.22), but this is equivalent to the asymptotic optimality). Non-asymptotic analogs of these properties also hold. (Our definition of the *Q-Bayes confidence predictor* is arbitrary in several respects: in principle, different choice functions can be used at different trials, the prediction can be arbitrary when $F(f(x)) = \max(\mathbf{M}(\epsilon), \mathbf{E}(\epsilon))$, and \mathbf{Y} can be replaced by any $E \subseteq \mathbf{Y}$ such that $Q(E | x) := \sum_{y \in E} Q(y | x) = 1$.)

The critical significance level (3.17) is an important characteristic of the probability distribution Q generating the individual examples. If $\epsilon > \epsilon_0$, the *Q-Bayes confidence predictor* will never output multiple predictions and, since it has to achieve the error rate ϵ , will sometimes have to output empty predictions. If, on the other hand, $\epsilon < \epsilon_0$, there will be multiple predictions but no empty predictions. Figures 3.2 and 3.3 suggest that the critical significance level for the permuted USPS data set is between 2% and 3%. This agrees with the observation that the critical significance level is just the error rate of the Bayes simple predictor (which is restricted to outputting prediction sets Γ_n with $|\Gamma_n| = 1$ and minimizes the expected number of errors) and the already mentioned fact (Vapnik 1998) that the error rate achieved by humans on the USPS data set is 2.5%. Notice that in Fig. 3.4 the onset of empty predictions closely follows the point where multiple predictions disappear; see also Figs. 3.10 and 3.11.

3.5 Proofs

First we establish some simple properties of the predictability distribution function and the multiplicity and emptiness curves.

Lemma 3.6. *The predictability distribution function F satisfies the following properties:*

1. $F(\epsilon) = 0$ for some $\epsilon > 0$ and $F(1) = 1$.
2. F is increasing.
3. F is continuous on the right.

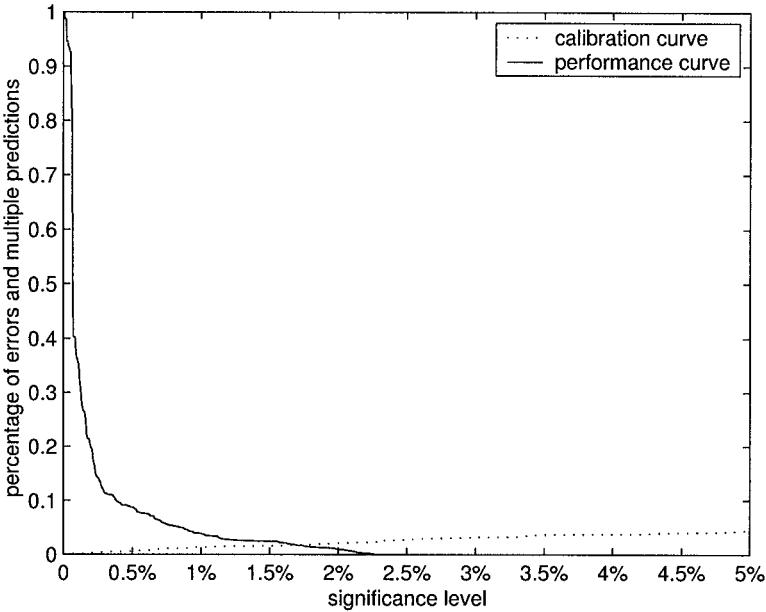


Fig. 3.10. Picture analogous to Fig. 3.6 for the last one thousand examples. Notice a different behavior of the empirical performance curve as it approaches the horizontal axis as compared with Fig. 3.6. The unexpected behavior of the empirical performance curve as it approaches the vertical axis may be explained by the presence of ambiguous and even misclassified examples (LeCun et al. 1990)

If a function $F : [0, 1] \rightarrow [0, 1]$ satisfies these properties, there exist a measurable space \mathbf{X} , a finite set \mathbf{Y} , and a probability distribution Q on $\mathbf{X} \times \mathbf{Y}$ for which F is the predictability distribution function.

Proof. Properties 1 (cf. the caption to Fig. 3.7), 2, and 3 are obvious (and the last two are well-known properties of all distribution functions). The fact that these three properties characterize predictability distribution functions easily follows from the fact that the last two properties plus $F(-\infty) = 0$ and $F(\infty) = 1$ characterize distribution functions (see, e.g., Shiryaev 1996, Theorem II.3.1). \square

We will use the notations g'_{left} and g'_{right} for the left and right derivatives, respectively, of a function g .

Lemma 3.7. *The multiplicity curve $M : [0, 1] \rightarrow [0, 1]$ always satisfies these properties:*

1. M is convex.
2. There is a point $\epsilon_0 \in [0, 1]$ (the critical significance level) such that $M(\epsilon) = 0$ for $\epsilon \geq \epsilon_0$ and $M'_{\text{left}}(\epsilon_0) < -1$; therefore, $M'_{\text{left}} < -1$ and $M'_{\text{right}} < -1$

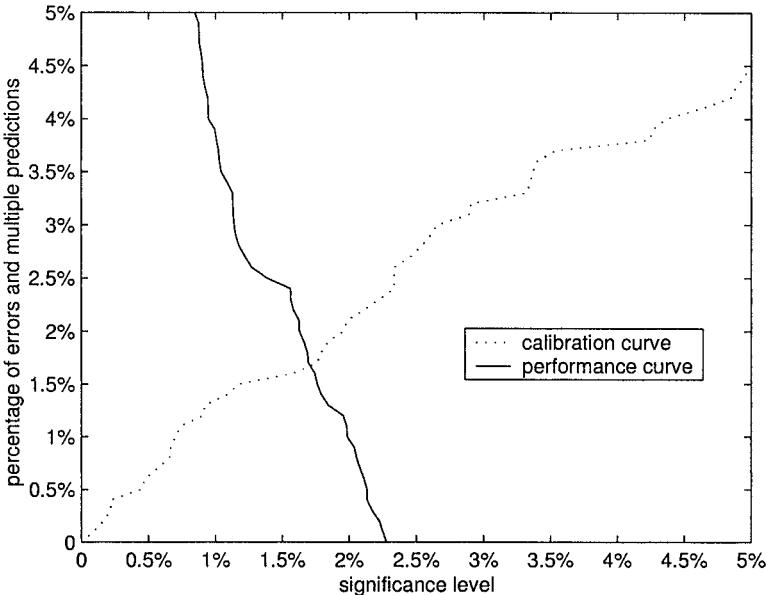


Fig. 3.11. The bottom part of Fig. 3.10 stretched vertically. Notice that the slope of the empirical performance curve is greater than 1 in absolute value before it hits the horizontal axis; this agrees with Lemma 3.7. This figure suggests that, if the 1-nearest neighbor conformal predictor were an optimal confidence predictor, the critical significance level for the permuted USPS data set would be close to 2.3%

to the left of ϵ_0 , and the function M is strictly decreasing before it hits the horizontal axis at ϵ_0 .

3. M is continuous at $\epsilon = 0$; therefore, it is continuous everywhere in $[0, 1]$.

If a function $M : [0, 1] \rightarrow [0, 1]$ satisfies these properties, there exist a measurable space X , a finite set Y , and a probability distribution Q on $X \times Y$ for which M is the multiplicity curve.

Proof sketch. For the basic properties of convex functions and their left and right derivatives, see, e.g., Bourbaki 1958 (§I.4). The statement of the lemma follows from the fact that the multiplicity curve M can be obtained from the predictability distribution function F using these steps (labeling the horizontal and vertical axes as x and y respectively):

1. Invert F : $F_1 := F^{-1}$.
2. Flip F_1 around the line $x = 0.5$ and then around the line $y = 0.5$: $F_2(x) := 1 - F_1(1 - x)$.
3. Integrate F_2 : $F_3(x) := \int_0^x F_2(t)dt$.
4. Invert F_3 : $F_4 := F_3^{-1}$.
5. Flip F_4 around the line $y = 0.5$: $F_5 := 1 - F_4$.

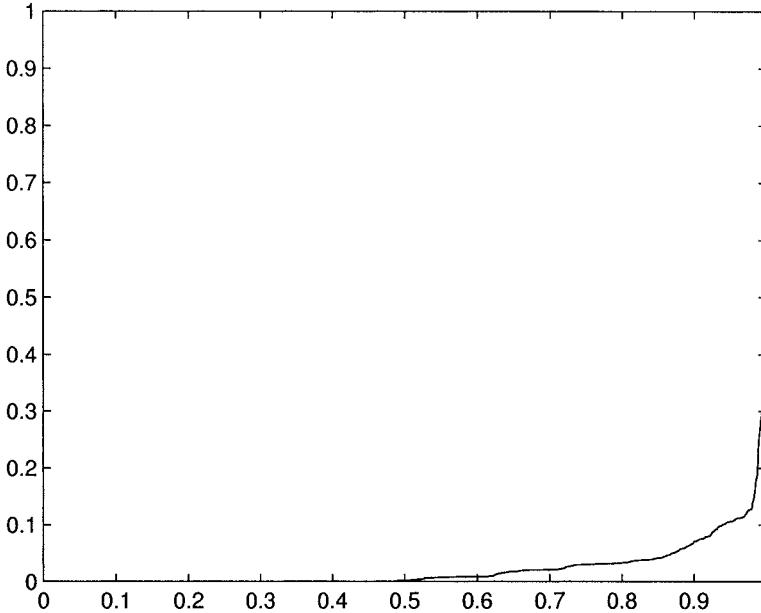


Fig. 3.12. An attempt to reverse engineer the predictability distribution function of the hand-written digits in the USPS data set. This picture was obtained from the solid line in Fig. 3.10 by reversing the list in the proof of Lemma 3.7

It can be shown that $\mathbf{M} = F_5$, no matter which of the several natural definitions of the operation $g \mapsto g^{-1}$ is used; for concreteness, we can define

$$g^{-1}(y) := \sup\{x : g(x) \leq y\} \quad (3.25)$$

for increasing g (so that g^{-1} is continuous on the right). \square

Propositions 3.3–3.5 suggest that if the 1-nearest neighbor conformal predictor is close to being optimal on the permuted USPS data set, its empirical performance curve is not far from the multiplicity curve \mathbf{M} . Visually the empirical performance curve in Figs. 3.5 and 3.6 seems to satisfy the properties listed in Lemma 3.7 for significance levels that are not too large or too small (approximately in the range 0.1%–5%); for an even better agreement, see Figs. 3.10 and 3.11.

A natural idea is to reverse the process of transforming F into \mathbf{M} and try to obtain an estimate of the predictability distribution function F from an empirical performance curve. Fig. 3.12 shows the result of such an attempt. Such pictures, however, should not be taken too seriously, since the differentiation operation needed in finding F is known to be unstable (see, e.g., Vapnik 1998, §1.12).

The following lemma parallels Lemma 3.7:

Lemma 3.8. *The emptiness curve $\mathbf{E} : [0, 1] \rightarrow [0, 1]$ always satisfies these properties:*

1. *There is a point $\epsilon_0 \in [0, 1]$ (namely, the critical significance level) such that $\mathbf{E}(\epsilon) = 0$ for $\epsilon \leq \epsilon_0$ and $\mathbf{E}(\epsilon)$ is concave for $\epsilon \geq \epsilon_0$.*
2. *$\mathbf{E}'_{\text{right}}(\epsilon_0) < \infty$ and $\mathbf{E}'_{\text{left}}(1) \geq 1$; therefore, for $\epsilon \in (\epsilon_0, 1)$, $1 \leq \mathbf{E}'_{\text{right}}(\epsilon) \leq \mathbf{E}'_{\text{left}}(\epsilon) < \infty$ and the function $\mathbf{E}(\epsilon)$ is strictly increasing.*
3. *$\mathbf{E}(\epsilon)$ is continuous at $\epsilon = \epsilon_0$; therefore, it is continuous everywhere in $[0, 1]$.*

If a function $\mathbf{E} : [0, 1] \rightarrow [0, 1]$ satisfies these properties, there exist a measurable space \mathbf{X} , a finite set \mathbf{Y} , and a probability distribution Q on $\mathbf{X} \times \mathbf{Y}$ for which \mathbf{E} is the emptiness curve.

Proof sketch. The statement of the lemma follows from the fact that the emptiness curve \mathbf{E} can be obtained from the predictability distribution function F using these steps:

1. Invert F : $F_1 := F^{-1}$.
2. Integrate F_1 : $F_2(x) := \int_0^x F_1(t)dt$.
3. Increase F_2 : $F_3(x) := F_2(x) + \epsilon_0$, where $\epsilon_0 := \int_0^1 F(x)dx$.
4. Invert F_3 : $F_4 := F_3^{-1}$.

It can be shown that $\mathbf{E} = F_4$, if we define $g^{-1}(y)$ by (3.25) (with $\sup \emptyset := 0$). \square

Proof of Proposition 3.2

Let $z_n = (x_n, y_n)$, $n = 1, 2, \dots$, be the examples output by Reality and τ_1, τ_2, \dots be the random numbers used by the nearest neighbors smoothed conformal predictor. Let w_1, w_2, \dots be the sequence of extended examples $w_n := (x_n, \tau'_n, y_n)$. Set

$$Q_n^\neq(y | x_i, \tau'_i) := \hat{Q}_{\{w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n\}}(y | x_i, \tau'_i), \quad (3.26)$$

for all $y \in \mathbf{Y}$ and $i = 1, \dots, n$. (The upper index \neq reminds us of the fact that (x_i, τ'_i) is not counted as one of its own nearest neighbors in this definition. For the definition of \hat{Q} , see (3.14) on p. 63.) We will also use the notation

$$f_n^\neq(x_i, \tau'_i) := \max_{y \in \mathbf{Y}} Q_n^\neq(y | x_i, \tau'_i) = \hat{f}_{\{w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n\}}(x_i, \tau'_i)$$

and let $\hat{y}_n(x_i, \tau'_i)$ (without the \neq , to make our notation less cumbersome) stand for the first element of $\arg \max_{y \in \mathbf{Y}} Q_n^\neq(y | x_i, \tau'_i)$ in a fixed ordering of \mathbf{Y} .

Without loss of generality we assume that $\{\epsilon_1, \dots, \epsilon_K\}$ contains only one significance level ϵ , which will be omitted from our notation. We will also assume that all extended objects $(x_i, \tau'_i) \in [0, 1]^2$ are different and that all

pairwise distances between them are also different (this is true with probability one, since τ'_i are independent random numbers uniformly distributed on $[0, 1]$). Our computational model has an operation of splitting $\tau \in [0, 1]$ into τ' and τ'' (or is allowed to generate both τ'_n and τ''_n at every trial n).

We will use two main data structures in our implementation of the nearest neighbors smoothed conformal predictor:

- a red-black binary *search tree* (see, e.g., Cormen et al. 2001, Chaps. 12–14; the only two operations on red-black trees we need in this book are the query **SEARCH** and the modifying operation **INSERT**);
- a growing *array* of nonnegative integers indexed by numbers $k \in \{-K_n, -K_n + 1, \dots, K_n\}$ (where n is the ordinal number of the example being processed).

Immediately after processing the n th extended example (x_n, τ_n, y_n) the contents of these data structures are as follows:

- The search tree contains n vertices, corresponding to the extended examples (x_i, τ_i, y_i) seen so far. The key of vertex i is the extended object $(x_i, \tau'_i) \in [0, 1]^2$; the linear order on the keys is the lexicographic order. The other information contained in vertex i is the random number τ''_i , the label y_i , the set $\{Q_n^{\neq}(y | x_i, \tau'_i) : y \in \mathbf{Y}\}$ of conditional probability estimates (3.26), the pointer to the following vertex (i.e., the vertex that has the smallest key greater than (x_i, τ'_i) ; if there is no greater key, the pointer is **NIL**), and the pointer to the previous vertex (i.e., the vertex that has the greatest key smaller than (x_i, τ'_i) ; if (x_i, τ'_i) is the smallest key, the pointer is **NIL**).
- The array contains the numbers

$$N(k) := |\{i = 1, \dots, n : \alpha_i = k/K_n\}| ,$$

with α_i defined by

$$\begin{aligned} \alpha_i := A &\left(\lceil(x_1, \tau'_1, y_1), \dots, (x_{i-1}, \tau'_{i-1}, y_{i-1}), \right. \\ &\left. (x_{i+1}, \tau'_{i+1}, y_{i+1}), \dots, (x_n, \tau'_n, y_n) \rceil, (x_i, \tau'_i, y_i) \right), \\ &i = 1, \dots, n , \quad (3.27) \end{aligned}$$

where the nonconformity measure A is defined by (3.15) on p. 63 (with $\hat{f}_B = f_n^{\neq}$ and $\hat{y}_B = \hat{y}_n$).

Notice that the information contained in vertex i of the search tree is sufficient to find $\hat{y}_n(x_i, \tau'_i)$ and α_i in time $O(1)$.

We will say that an extended object (x_j, τ'_j) is in the *vicinity* of an extended object (x_i, τ'_i) if there are less than K_n extended objects (x_k, τ'_k) (strictly) between (x_i, τ'_i) and (x_j, τ'_j) , in the sense of the lexicographic order.

When a new object x_n becomes known, the algorithm does the following:

- Generates τ'_n and τ''_n .
- Locates the successor and predecessor of (x_n, τ'_n) in the search tree (using the query SEARCH and the pointers to the following and previous vertices); this requires time $O(\log n)$.
- Computes the estimated conditional probabilities $\{Q_n^\neq(y | x_n, \tau'_n) : y \in \mathbf{Y}\}$; this also gives $\hat{y}_n(x_n, \tau'_n)$. This involves scanning the vicinity of (x_n, τ'_n) for the K_n nearest neighbors of (x_n, τ'_n) , which can be done in time $O(K_n)$: the K_n nearest neighbors can be extracted from the vicinity of (x_n, τ'_n) sorted in the order of increasing distances from (x_n, τ'_n) ; since initially the vicinity consists of two sorted lists (to the left and to the right of (x_n, τ'_n)), the procedure MERGE used in the merge sort algorithm (see, e.g., Cormen et al. 2001, §2.3.1) will sort the whole vicinity in time $O(K_n)$. Therefore, the required time is $O(K_n) = O(\log n)$.
- For each $y \in \mathbf{Y}$ looks at what happens if the n th example is $(x_n, \tau_n, y_n) = (x_n, \tau_n, y)$: computes α_n and updates (if necessary) α_i for (x_i, τ'_i) in the vicinity of (x_n, τ'_n) ; using the array and τ''_n , finds whether $y \in \Gamma_n$. This requires time $O(K_n^2) = O(\log n)$, since there are $O(K_n)$ α_i 's in the vicinity of (x_n, τ'_n) and each of them can be computed in time $O(K_n)$.
- Outputs the prediction set Γ_n (time $O(1)$).

When the label y_n arrives, the algorithm:

- Inserts the new vertex $(x_n, \tau'_n, \tau''_n, y_n, \{Q_n^\neq(y | x_n, \tau'_n) : y \in \mathbf{Y}\})$ in the search tree, repairs the pointers to the following and previous elements for (x_n, τ'_n) 's left and right neighbors, initializes the pointers to the following and previous elements for (x_n, τ'_n) itself, and rebalances the tree (time $O(\log n)$).
- Updates (if necessary) the conditional probabilities

$$\{Q_{n-1}^\neq(y | x_i, \tau'_i) : y \in \mathbf{Y}\} \mapsto \{Q_n^\neq(y | x_i, \tau'_i) : y \in \mathbf{Y}\}$$

for the $2K_n$ existing vertices (x_i, τ'_i) in the vicinity of (x_n, τ'_n) ; this requires time $O(K_n^2) = O(\log n)$. The conditional probabilities for the other (x_i, τ'_i) , $i = 1, \dots, n-1$, do not change.

- Updates the array, changing $N(K_n \alpha_i)$ for the $(x_i, \tau'_i) \neq (x_n, \tau'_n)$ in the vicinity of (x_n, τ'_n) and for both old and new values of α_i and changing $N(K_n \alpha_n)$ (time $O(K_n) = O(\log n)$).

In conclusion we discuss how to do the updates required when K_n changes. At the critical trials n when K_n changes the array and all estimated conditional probabilities $Q_n^\neq(y | x_i, \tau'_i)$ have to be recomputed, which, if done naively, would require time $\Theta(n K_n)$.

The assumption we have made about K_n so far is that $K_n = O(\sqrt{\log n})$. Now we also assume that K_n is strictly increasing and

$$|\{n : K_n < c\}| = O(|\{n : K_n = c\}|) \tag{3.28}$$

as $c \rightarrow \infty$. This is the full explication of the “ $K_n \rightarrow \infty$ sufficiently slowly” in the statement of the lemma, as used in this proof.

An *epoch* is defined to be a maximal sequence of ns with the same K_n . Since the changes that need to be done when a new epoch starts are substantial, they will be spread over the whole preceding epoch. An epoch is *odd* if the corresponding K_n is odd and *even* if K_n is even. At every trial in an epoch we prepare the ground for the next epoch. We will only discuss updating the estimated conditional probabilities $Q_n^\#(y | x_i, \tau'_i)$; the array is treated in a similar way.

By the end of epoch $n = A + 1, A + 2, \dots, B$ we need to change B sets $\{Q_n^\#(y | x_i, \tau'_i) : y \in \mathbf{Y}\}$ in $B - A$ trials (the duration of the epoch). Therefore, each vertex of the search tree should contain not only $\{Q_n^\#(y | x_i, \tau'_i)\}$ for the current epoch but also $\{Q_n^\#(y | x_i, \tau'_i)\}$ for the next epoch (two structures for holding $\{Q_n^\#(y | x_i, \tau'_i)\}$ will suffice, one for even epochs and one for odd epochs). Our assumptions of the slow growth of K_n (see (3.28)) imply that $B = O(B - A)$. This means that at each trial $O(1)$ sets $\{Q_n^\#(y | x_i, \tau'_i)\}$ for the next epoch should be added. This will take time $O(K_n) = O(\log n)$. As soon as a set $\{Q_n^\#(y | x_i, \tau'_i) : y \in \mathbf{Y}\}$ for the next epoch is added at some trial, both sets (for the current and next epoch) will have to be updated for each new example.

Proof of Proposition 3.3

Let us check first that (3.20) indeed implies $\mathbf{P}(\epsilon) \geq \mathbf{M}(\epsilon)$ (we will omit the lower indices Γ, Q). Since probability distributions are σ -additive, (3.19) implies

$$\limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \leq \mathbf{P}(\epsilon) \quad \text{a.s. ,}$$

and so we obtain from (3.20):

$$\mathbf{P}(\epsilon) \geq \limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \geq \liminf_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma, Q^\infty)}{n} \geq \mathbf{M}(\epsilon)$$

almost surely; since the two extreme terms are deterministic, we have $\mathbf{P}(\epsilon) \geq \mathbf{M}(\epsilon)$.

We start the actual proof with alternative definitions of calibration and performance curves. Complement the basic protocol given at the beginning of this chapter (p. 53) in which Reality plays Q^∞ and Predictor plays Γ with the following variables:

$$\begin{aligned} \overline{\text{err}}_n^\epsilon := (Q \times \mathbf{U}) &\left\{ (x, y, \tau) \in (\mathbf{Z} \times [0, 1]) : \right. \\ &\left. y \notin \Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau) \right\} , \end{aligned} \quad (3.29)$$

$$\overline{\text{mult}}_n^\epsilon := (Q_{\mathbf{X}} \times \mathbf{U}) \left\{ (x, \tau) \in (\mathbf{X} \times [0, 1]) : |\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)| > 1 \right\}, \quad (3.30)$$

$$\overline{\text{Err}}_n^\epsilon := \sum_{i=1}^n \overline{\text{err}}_i^\epsilon, \quad \overline{\text{Mult}}_n^\epsilon := \sum_{i=1}^n \overline{\text{mult}}_i^\epsilon. \quad (3.31)$$

The *predictable calibration curve* of Γ under Q is defined by

$$\overline{\mathbf{C}}(\epsilon) := \inf \left\{ \beta : \mathbb{P} \left\{ \limsup_{n \rightarrow \infty} \frac{\overline{\text{Err}}_n^\epsilon}{n} \leq \beta \right\} = 1 \right\}$$

and the *predictable performance curve* of Γ under Q by

$$\overline{\mathbf{P}}(\epsilon) := \inf \left\{ \beta : \mathbb{P} \left\{ \limsup_{n \rightarrow \infty} \frac{\overline{\text{Mult}}_n^\epsilon}{n} \leq \beta \right\} = 1 \right\},$$

where \mathbb{P} refers to the probability distribution $(Q \times \mathbf{U})^\infty$ over the examples z_1, z_2, \dots and random numbers τ_1, τ_2, \dots . By the martingale strong law of large numbers (see §A.6) the predictable versions of the calibration and performance curves coincide with the original versions: indeed, since $\text{Err}_n^\epsilon - \overline{\text{Err}}_n^\epsilon$ and $\text{Mult}_n^\epsilon - \overline{\text{Mult}}_n^\epsilon$ are martingales (with increments bounded by 1 in absolute value) for all ϵ with respect to the filtration \mathcal{F}_n , $n = 0, 1, \dots$, where each \mathcal{F}_n is generated by z_1, \dots, z_n and τ_1, \dots, τ_n , we have

$$\lim_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon - \overline{\text{Err}}_n^\epsilon}{n} = 0 \quad \mathbb{P}\text{-a.s.}$$

and

$$\lim_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon - \overline{\text{Mult}}_n^\epsilon}{n} = 0 \quad \mathbb{P}\text{-a.s.}$$

It is also clear that we can replace Mult_n^ϵ by $\overline{\text{Mult}}_n^\epsilon$ in (3.20).

Without loss of generality we can assume that Predictor's move Γ_n^ϵ at trial n is, for each ϵ , either $\{\hat{y}(x_n)\}$ (\hat{y} is defined by (3.24), p. 70) or vacuous, the whole label space \mathbf{Y} . Furthermore, we can assume that

$$\text{mult}_n^\epsilon = \mathbf{M}(\overline{\text{err}}_n^\epsilon) \quad (3.32)$$

at every trial, since the best way to spend the allowance of $\overline{\text{err}}_n^\epsilon$ is to give non-vacuous predictions for objects x with the largest (topmost in Figs. 3.7 and 3.8) representations $F(f(x))$. (For a formal argument, see the end of this proof.) Using the fact that the multiplicity curve \mathbf{M} is convex, decreasing, and continuous (see Lemma 3.7), we obtain, for any significance level ϵ ,

$$\begin{aligned} \frac{\overline{\text{Mult}}_n^\epsilon}{n} &= \frac{1}{n} \sum_{i=1}^n \overline{\text{mult}}_i^\epsilon = \frac{1}{n} \sum_{i=1}^n \mathbf{M}(\overline{\text{err}}_i^\epsilon) \\ &\geq \mathbf{M} \left(\frac{1}{n} \sum_{i=1}^n \overline{\text{err}}_i^\epsilon \right) = \mathbf{M} \left(\frac{\overline{\text{Err}}_n^\epsilon}{n} \right) \geq \mathbf{M}(\epsilon) - \delta, \end{aligned} \quad (3.33)$$

the last inequality holding almost surely for an arbitrary $\delta > 0$ from some n on.

It remains to prove formally that $\overline{\text{mult}}_n^\epsilon \geq M(\overline{\text{err}}_n^\epsilon)$ (which is the part of (3.32) that we actually used). Let us fix ϵ and

$$x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1};$$

we will write

$$\Gamma(x, \tau) := \Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau),$$

omitting the fixed arguments. Without loss of generality we are assuming that either $\Gamma(x, \tau) = \{\hat{y}(x)\}$ or $\Gamma(x, \tau) = \mathbf{Y}$. Set

$$p(x) := \mathbf{U}\{\tau : \Gamma(x, \tau) = \{\hat{y}(x)\}\}, \quad \delta := \overline{\text{err}}_n.$$

Our goal is to show that $\overline{\text{mult}}_n \geq M(\delta)$; without loss of generality we assume $\delta < \epsilon_0$, where ϵ_0 is the critical significance level. To put it differently, we are required to show that the value of the optimization problem

$$\int_{\mathbf{X}} p(x) Q(dx) \rightarrow \max \tag{3.34}$$

subject to the constraint

$$\int_{\mathbf{X}} (1 - f(x)) p(x) Q(dx) = \delta$$

is $1 - M(\delta)$ at best (remember that $f(x)$ is the predictability of x ; Q is a shorthand for $Q_{\mathbf{X}}$). By the Neyman–Pearson lemma (see, e.g., Lehmann 1986, Theorem 3.2.1) for some solution p there exist constants $c > 0$ and $d \in [0, 1]$ such that

$$p(x) = \begin{cases} 1 & \text{if } f(x) > c \\ d & \text{if } f(x) = c \\ 0 & \text{if } f(x) < c. \end{cases} \tag{3.35}$$

The constants c and d are defined (c uniquely and d uniquely unless the probability of $f(x) = c$ is zero or $c = 1$; in the latter case, $d = 1$) from the condition

$$\int_{x: f(x) > c} (1 - f(x)) Q(dx) + d \int_{x: f(x) = c} (1 - c) Q(dx) = \delta,$$

which is equivalent (see Fig. 3.7 on p. 66) to

$$\int_0^1 (F(\beta) - F(c))^+ d\beta + d(1 - c)(F(c) - F(c^-)) = \delta, \tag{3.36}$$

where $F(c-)$ is defined as $\lim_{\beta \uparrow c} F(\beta)$. From this it is easy to obtain that the value of the optimization problem (3.34) is indeed $1 - \mathbf{M}(\delta)$: using the notation $p_d(x)$ for the right-hand side of (3.35), we have

$$\begin{aligned}\int_{\mathbf{X}} p_d(x) Q(dx) &= d \int p_1(x) Q(dx) + (1-d) \int p_0(x) Q(dx) \\ &= dQ\{x : f(x) \geq c\} + (1-d)Q\{x : f(x) > c\} \\ &= d(1 - F(c-)) + (1-d)(1 - F(c)) \\ &= 1 - F(c) + d(F(c) - F(c-)) = 1 - \mathbf{M}(\delta),\end{aligned}$$

the last equality following from (3.36) (except for the case $c = 1$, when it is obvious). This completes the proof.

Proof sketch of Proposition 3.4

The proof of Proposition 3.4 is similar to (but more complicated than) the proof of Proposition 3.3; this sketch can be made rigorous using the Neyman–Pearson lemma, as we did in the proof of Proposition 3.3.

Along with the random variables (3.29)–(3.31) we will also need

$$\overline{\text{emp}}_n^\epsilon := (Q_{\mathbf{X}} \times \mathbf{U}) \left\{ (x, \tau) \in (\mathbf{X} \times [0, 1]) : |\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)| = 0 \right\} \quad (3.37)$$

and

$$\overline{\text{Emp}}_n^\epsilon := \sum_{i=1}^n \overline{\text{emp}}_i^\epsilon. \quad (3.38)$$

It is clear that

$$\lim_{n \rightarrow \infty} \frac{\overline{\text{Emp}}_n^\epsilon - \overline{\text{Emp}}_n^\epsilon}{n} = 0 \quad \text{a.s.}$$

Without loss of generality we can assume that Predictor's move Γ_n at trial n is $\{\hat{y}(x_n)\}$ or the empty set \emptyset or the whole label space \mathbf{Y} . Furthermore, we can assume that, at every trial, the predictions are singular (i.e., contain one label) for the new objects above the straight line BC in Fig. 3.13 (more formally, for new extended objects (x, τ) satisfying

$$F(x, \tau) := F(f(x)-) + \tau (F(f(x)+) - F(f(x)-)) \geq \mathbf{M}(\overline{\text{err}}_n^\epsilon - \overline{\text{emp}}_n^\epsilon);$$

intuitively, considering extended objects makes the vertical axis “infinitely divisible”) and that the predictions are empty for the objects below the straight line DG in Fig. 3.13. (Indeed, predictions of this kind are admissible in the sense that we cannot improve $\overline{\text{mult}}_n^\epsilon$ and $\overline{\text{emp}}_n^\epsilon$ simultaneously, and all admissible predictions are equivalent to predictions of this kind. A formal argument for the case where emp_n^ϵ are omitted is given in the proof of Proposition 3.3 above.) It is clear that for the confidence predictor to satisfy (3.21) it must hold that

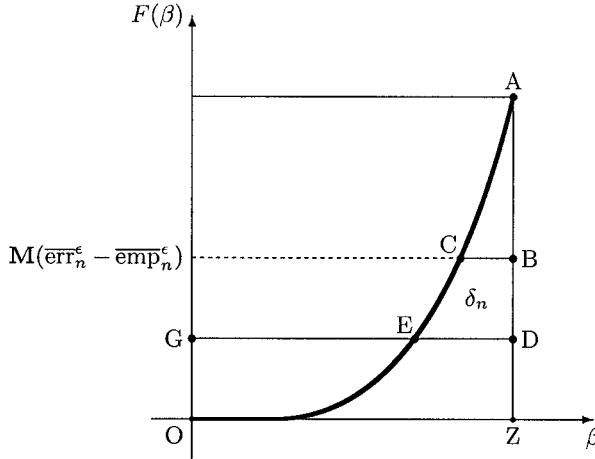


Fig. 3.13. An admissible confidence predictor. The thick line is the predictability distribution function F ; the area of the curvilinear triangle ABC is $\overline{\text{err}}_n^\epsilon - \overline{\text{emp}}_n^\epsilon$; the area of the rectangle $DZOG$ is $\overline{\text{emp}}_n^\epsilon$; the (nonnegative) area of the curvilinear quadrangle $BDEC$ is denoted δ_n

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\delta_i \wedge \overline{\text{emp}}_i^\epsilon) = 0$$

(otherwise $\overline{\text{Mult}}_n^\epsilon$ can be decreased substantially, which contradicts (3.20); δ_i are defined in the caption of Fig. 3.13), and so we can assume, without loss of generality, that either $\delta_n = 0$ or $\overline{\text{emp}}_n^\epsilon = 0$ at every trial n , i.e., that

$$\overline{\text{mult}}_n^\epsilon = \mathbf{M}(\overline{\text{err}}_n^\epsilon), \quad \overline{\text{emp}}_n^\epsilon = \mathbf{E}(\overline{\text{err}}_n^\epsilon)$$

at every trial. In the sequel we will omit the upper index ϵ .

Let us check that to achieve (3.21) the randomized confidence predictor must satisfy

$$\epsilon < \epsilon_0 \implies \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\overline{\text{err}}_i - \epsilon_0)^+ = 0 \quad (3.39)$$

$$\epsilon \geq \epsilon_0 \implies \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\epsilon_0 - \overline{\text{err}}_i)^+ = 0, \quad (3.40)$$

where the convergence is, as usual, almost certain. We know from Lemma 3.7 that the multiplicity curve M is convex, decreasing, continuous, and has slope at most -1 before it hits the horizontal axis at $\epsilon = \epsilon_0$. The second implication, (3.40), now immediately follows from the fact that, under $\epsilon \geq \epsilon_0$ and (3.21),

$$0 = \limsup_{n \rightarrow \infty} \frac{\overline{\text{Mult}}_n}{n} = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{M}(\overline{\text{err}}_i) \geq \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\epsilon_0 - \overline{\text{err}}_i)^+$$

The first implication, (3.39), can be extracted from the chain (3.33) in the proof of Proposition 3.3. Indeed, it can be seen from (3.33) that, assuming the predictor satisfies (3.21) and $\epsilon < \epsilon_0$,

$$\overline{\text{Err}}_n / n \rightarrow \epsilon \quad \text{a.s.}$$

and, therefore,

$$\begin{aligned} \mathbf{M}(\epsilon) &\geq \limsup_{n \rightarrow \infty} \frac{\overline{\text{Mult}}_n}{n} = \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{M}(\overline{\text{err}}_i) \\ &= \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{M}(\overline{\text{err}}_i \wedge \epsilon_0) \geq \limsup_{n \rightarrow \infty} \mathbf{M} \left(\frac{1}{n} \sum_{i=1}^n (\overline{\text{err}}_i \wedge \epsilon_0)^+ \right) \\ &= \limsup_{n \rightarrow \infty} \mathbf{M} \left(\frac{\overline{\text{Err}}_n}{n} - \frac{1}{n} \sum_{i=1}^n (\overline{\text{err}}_i - \epsilon_0)^+ \right) \\ &= \limsup_{n \rightarrow \infty} \mathbf{M} \left(\epsilon - \frac{1}{n} \sum_{i=1}^n (\overline{\text{err}}_i - \epsilon_0)^+ \right) \\ &= \mathbf{M} \left(\epsilon - \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (\overline{\text{err}}_i - \epsilon_0)^+ \right) \end{aligned}$$

almost surely. This proves (3.39).

Using (3.39), (3.40), and the fact that the emptiness curve \mathbf{E} is concave, increasing, and (uniformly) continuous for $\epsilon \geq \epsilon_0$ (see Lemma 3.8), we obtain: if $\epsilon < \epsilon_0$,

$$\begin{aligned} \frac{\overline{\text{Emp}}_n}{n} &= \frac{1}{n} \sum_{i=1}^n \overline{\text{emp}}_i = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(\overline{\text{err}}_i) \\ &\leq \frac{1}{n} \mathbf{E}'_{\text{right}}(\epsilon_0) \sum_{i=1}^n (\overline{\text{err}}_i - \epsilon_0)^+ \rightarrow 0 \quad (n \rightarrow \infty); \end{aligned}$$

if $\epsilon \geq \epsilon_0$,

$$\begin{aligned} \frac{\overline{\text{Emp}}_n}{n} &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}(\overline{\text{err}}_i) = \frac{1}{n} \sum_{i=1}^n \mathbf{E}(\overline{\text{err}}_i \vee \epsilon_0) \\ &\leq \mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n (\overline{\text{err}}_i \vee \epsilon_0)^+ \right) = \mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n \overline{\text{err}}_i + \frac{1}{n} \sum_{i=1}^n (\epsilon_0 - \overline{\text{err}}_i)^+ \right) \\ &= \mathbf{E} \left(\frac{1}{n} \sum_{i=1}^n \overline{\text{err}}_i \right) + o(1) \leq \mathbf{E}(\epsilon) + \delta, \end{aligned}$$

the last inequality holding almost surely for an arbitrary $\delta > 0$ from some n on and ϵ being the significance level used.

Proof sketch of Proposition 3.5

It will be convenient to consider a modification and extension of the function $Q_n^{\neq}(y | x_i, \tau'_i)$ introduced in (3.26). An alternative definition of the nearest neighbors approximations $Q_n(y | x, \sigma)$ to the conditional probabilities $Q(y | x)$ is as follows: for every $(x, \sigma, y) \in \tilde{\mathbf{Z}}$,

$$Q_n(y | x, \sigma) := \hat{Q}_{\{w_1, \dots, w_n\}}(y | x, \sigma).$$

(This time (x_i, τ'_i) is not prevented from being counted as one of the K_n nearest neighbors of (x, σ) if $(x_i, \tau'_i) = (x, \sigma)$.) We define the empirical predictability function f_n by

$$f_n(x, \sigma) := \max_{y \in \mathbf{Y}} Q_n(y | x, \sigma) = \hat{f}_{\{w_1, \dots, w_n\}}(x, \sigma).$$

The proof will be based on the following version of a well-known fundamental result.

Lemma 3.9. *Suppose $K_n \rightarrow \infty$, $K_n = o(n)$, and $\mathbf{Y} = \{0, 1\}$. For any $\delta > 0$ and large enough n ,*

$$\mathbb{P} \left\{ \int |Q(1 | x) - Q_n(1 | x, \sigma)| Q_{\mathbf{X}}(dx) U(d\sigma) > \delta \right\} \leq e^{-n\delta^2/40},$$

where the outermost probability distribution \mathbb{P} (essentially $(Q \times U)^\infty$) generates the extended examples (x_i, τ_i, y_i) , which determine the empirical distributions Q_n .

Proof. This is almost a special case of Devroye et al.'s (1994) Theorem 1. There is, however, an important difference between the way we break distance ties and the way Devroye et al. (1994) do this: in Devroye et al. 1994, instead of our (3.10),

$$(|x_1 - x_3|, |\sigma_1 - \sigma_3|) < (|x_2 - x_3|, |\sigma_2 - \sigma_3|)$$

is used. (Our way of breaking ties better agrees with the lexicographic order on $[0, 1]^2$, which is useful in the proof of Proposition 3.2 and, less importantly, in the proof of Lemma 3.11.) It is easy to check that the proof given in Devroye et al. 1994 also works (and becomes simpler) for our way of breaking distance ties. \square

Lemma 3.10. *Suppose $K_n \rightarrow \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n ,*

$$\mathbb{P} \left\{ (Q_{\mathbf{X}} \times U) \left\{ (x, \sigma) : \max_{y \in \mathbf{Y}} |Q(y | x) - Q_n(y | x, \sigma)| > \delta \right\} > \delta \right\} \leq e^{-\delta^* n};$$

in particular,

$$\mathbb{P} \{ (Q_{\mathbf{X}} \times U) \{ (x, \sigma) : |f(x) - f_n(x, \sigma)| > \delta \} > \delta \} \leq e^{-\delta^* n}.$$

Proof. We apply Lemma 3.9 to the binary classification problem obtained from our classification problem by replacing label $y \in \mathbf{Y}$ with 1 and replacing all other labels with 0:

$$\mathbb{P} \left\{ \int |Q(y|x) - Q_n(y|x, \sigma)| Q_{\mathbf{X}}(dx) \mathbf{U}(d\sigma) > \delta \right\} \leq e^{-n\delta^2/40}.$$

By Markov's inequality this implies

$$\mathbb{P} \left\{ (Q_{\mathbf{X}} \times \mathbf{U}) \{ |Q(y|x) - Q_n(y|x, \sigma)| > \sqrt{\delta} \} > \sqrt{\delta} \right\} \leq e^{-n\delta^2/40},$$

which, in turn, implies

$$\mathbb{P} \left\{ (Q_{\mathbf{X}} \times \mathbf{U}) \left\{ \max_{y \in \mathbf{Y}} |Q(y|x) - Q_n(y|x, \sigma)| > \sqrt{\delta} \right\} > |\mathbf{Y}| \sqrt{\delta} \right\} \leq e^{-n\delta^2/40}.$$

This completes the proof, since we can take the δ in the last equation arbitrarily small as compared to the δ in the statement of the lemma. \square

We will use the shorthand “ $\forall^\infty n$ ” for “from some n on”.

Lemma 3.11. *Suppose $K_n \rightarrow \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n ,*

$$\mathbb{P} \left\{ \frac{|\{i : \max_y |Q(y|x_i) - Q_n^{\neq}(y|x_i, \tau'_i)| > \delta\}|}{n} > \delta \right\} \leq e^{-\delta^* n}.$$

In particular,

$$\forall^\infty n : \mathbb{P} \left\{ \frac{|\{i : |f(x_i) - f_n^{\neq}(x_i, \tau'_i)| > \delta\}|}{n} > \delta \right\} \leq e^{-\delta^* n}.$$

Proof. Since

$$|Q_n^{\neq}(y|x_i, \tau'_i) - Q_n(y|x_i, \tau'_i)| \leq \frac{1}{K_n} = o(1),$$

we can, and will, ignore the upper indices \neq in the statement of the lemma.

Define

$$I_n(x, \sigma) := \begin{cases} 0 & \text{if } \max_y |Q(y|x) - Q_n(y|x, \sigma)| \leq \delta \\ 1 & \text{if } \max_y |Q(y|x) - Q_n(y|x, \sigma)| \geq 2\delta \\ (\max_y |Q(y|x) - Q_n(y|x, \sigma)| - \delta)/\delta & \text{otherwise} \end{cases}$$

(intuitively, $I_n(x, \sigma)$ is a “soft version” of $\mathbb{I}_{\max_y |Q(y|x) - Q_n(y|x, \sigma)| > \delta}$).

The main tool in this proof (and several other proofs in this section) will be McDiarmid's theorem (see §A.7). First we check the possibility of its application. If we replace an extended object (x_j, τ'_j) by another extended object (x_j^*, τ_j^*) , the expression

$$\sum_{i=1}^n I_n(x_i, \tau'_i)$$

will change as follows:

- the addend $I_n(x_i, \tau'_i)$ for $i = j$ changes by 1 at most;
- the addends $I_n(x_i, \tau'_i)$ for $i \neq j$ such that neither (x_j, τ'_j) nor (x_j^*, τ_j^*) are among the K_n nearest neighbors of (x_i, τ'_i) do not change at all;
- the sum over the at most $4K_n$ (see below) addends $I_n(x_i, \tau'_i)$ for $i \neq j$ such that either (x_j, τ'_j) or (x_j^*, τ_j^*) (or both) are among the K_n nearest neighbors of (x_i, τ'_i) can change by at most

$$4K_n \frac{1}{\delta} \frac{1}{K_n} = \frac{4}{\delta}. \quad (3.41)$$

The left-hand side of (3.41) reflects the following facts: the change in $Q_n(y | x_i, \tau'_i)$ for $i \neq j$ is at most $1/K_n$; the number of $i \neq j$ such that (x_j, τ'_j) is among the K_n nearest neighbors of (x_i, τ'_i) does not exceed $2K_n$ (since the extended objects are linearly ordered and (3.10) is used for breaking distance ties); analogously, the number of $i \neq j$ such that (x_j^*, τ_j^*) is among the K_n nearest neighbors of (x_i, τ'_i) does not exceed $2K_n$.

Therefore, by McDiarmid's theorem,

$$\begin{aligned} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n I_n(x_i, \tau'_i) - \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n I_n(x_i, \tau'_i) \right) > \delta \right\} \\ \leq \exp \left(-2\delta^2 n / (1 + 4/\delta)^2 \right) = \exp \left(-\frac{2\delta^4}{(4 + \delta)^2} n \right). \end{aligned} \quad (3.42)$$

Next we find:

$$\begin{aligned} \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n I_n(x_i, \tau'_i) \right) &= \mathbb{E} (I_n(x_n, \tau'_n)) \leq \mathbb{E} (I_{n-1}(x_n, \tau'_n)) + o(1) \\ &\leq \mathbb{E}(Q_{\mathbf{X}} \times \mathbf{U}) \{ (x, \sigma) : \max_y |Q(y | x) - Q_{n-1}(y | x, \sigma)| > \delta \} + o(1) \\ &\leq e^{-\delta^*(n-1)} + \delta + o(1) \leq 2\delta \end{aligned}$$

(the penultimate inequality follows from Lemma 3.10) from some n on. In combination with (3.42) this implies

$$\forall^\infty n : \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n I_n(x_i, \tau'_i) > 3\delta \right\} \leq \exp \left(-\frac{2\delta^4}{(4 + \delta)^2} n \right),$$

in particular

$$\mathbb{P} \left\{ \frac{|\{i : \max_y |Q(y | x_i) - Q_n(y | x_i, \tau'_i)| \geq 2\delta\}|}{n} > 3\delta \right\} \leq \exp \left(-\frac{2\delta^4}{(4 + \delta)^2} n \right).$$

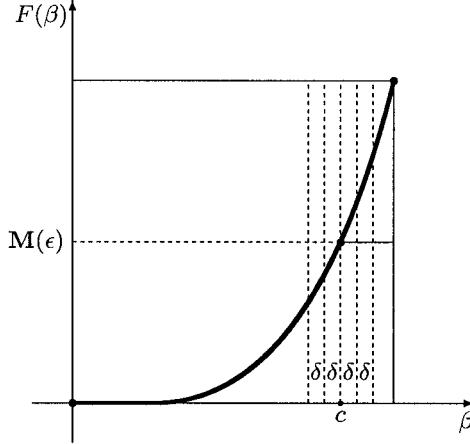


Fig. 3.14. Case $F(c) = M(\epsilon)$

Replacing 3δ by δ , we obtain that, from some n on,

$$\mathbb{P} \left\{ \frac{|\{i : \max_y |Q(y | x_i) - Q_n(y | x_i, \tau'_i)| > \delta\}|}{n} > \delta \right\} \leq \exp \left(-\frac{2(\delta/3)^4}{(4 + \delta/3)^2 n} \right),$$

which completes the proof. \square

We say that an extended example (x_i, τ_i, y_i) , $i = 1, \dots, n$, is n -strange if $y_i \neq \hat{y}_n(x_i, \tau'_i)$; otherwise, (x_i, τ_i, y_i) will be called n -conforming. We will assume that $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$, $i = 1, \dots, n$, are all different for all n ; even more than that, we will assume that τ''_i , $i = 1, 2, \dots$, are all different (we can do so since the probability of this event is one).

Lemma 3.12. Suppose (3.13) (p. 63) is satisfied and $\epsilon \leq \epsilon_0$. With probability one, the $\lfloor (1 - M(\epsilon))n \rfloor$ extended examples with the largest (in the sense of the lexicographic order) $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$ among $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$ contain at most $n\epsilon + o(n)$ n -strange extended examples as $n \rightarrow \infty$.

Proof. Define

$$c := \sup \{ \beta : F(\beta) \leq M(\epsilon) \}.$$

It is clear that $0 < c < 1$. Our proof will work both in the case where $F(c) = M(\epsilon)$ and in the case where $F(c) > M(\epsilon)$, as illustrated in Figs. 3.14 and 3.15.

Let $\delta > 0$ be a small constant (we will let $\delta \rightarrow 0$ eventually). Define a “threshold” $(c'_n, c''_n) \in [0, 1]^2$ requiring that

$$\mathbb{P} \{ f(x_n) = c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n) \} = F(c) - M(\epsilon) - \delta \quad (3.43)$$

if $F(c) > M(\epsilon)$; we assume that δ is small enough for

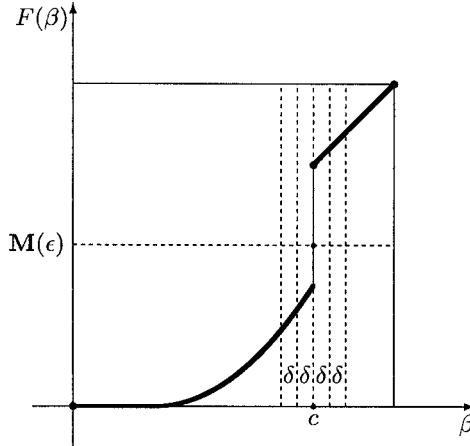


Fig. 3.15. Case $F(c) > M(\epsilon)$

$$2\delta < F(c) - M(\epsilon) \quad (3.44)$$

to hold (among other things this will ensure the validity of the definition (3.43)). If $F(c) = M(\epsilon)$, we set $(c'_n, c''_n) := (c + \delta, 0)$; in any case, we will have

$$\mathbb{P}\{f(x_n) = c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n)\} \geq F(c) - M(\epsilon) - \delta. \quad (3.45)$$

Let us say that an extended example (x_i, τ_i, y_i) is *above the threshold* if

$$(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i) > (c'_n, c''_n);$$

otherwise, we say it is *below the threshold*. Divide the first n extended examples (x_i, τ_i, y_i) , $i = 1, \dots, n$, into five classes:

Class I: Those satisfying $f(x_i) \leq c - 2\delta$.

Class II: Those that satisfy $f(x_i) = c$ and are below the threshold.

Class III: Those satisfying $c - 2\delta < f(x_i) \leq c + 2\delta$ but not $f(x_i) = c$.

Class IV: Those that satisfy $f(x_i) = c$ and are above the threshold.

Class V: Those satisfying $f(x_i) > c + 2\delta$.

First we explain the general idea of the proof. The threshold (c'_n, c''_n) was chosen so that approximately $\lfloor (1 - M(\epsilon))n \rfloor$ of the available extended examples will be above the threshold. Because of this, the extended examples above the threshold will essentially be the $\lfloor (1 - M(\epsilon))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$ referred to in the statement of the lemma. For each of the five classes we will be interested in the following questions:

- How many extended examples are there in the class?

- How many of those are above the threshold?
- How many of those above the threshold are n -strange?

If the sum of the answers to the last question does not exceed $n\epsilon$ by too much, we are done.

With this plan in mind, we start the formal proof. (Of course, we will not be following the plan literally: for example, if a class is very small, we do not need to answer the second and third questions.) The first step is to show that

$$c - \delta \leq c'_n \leq c + \delta \quad (3.46)$$

from some n on; this will ensure that the classes are conveniently separated from each other. We only need to consider the case $F(c) > \mathbf{M}(\epsilon)$. The inequality $c'_n \leq c + \delta$ follows from

$$\forall^\infty n : \mathbb{P}\{f(x_n) = c, f_{n-1}(x_n, \tau'_n) > c + \delta\} < \delta < F(c) - \mathbf{M}(\epsilon) - \delta$$

(combine Lemma 3.10 with (3.44)). The inequality $c - \delta \leq c'_n$ follows from

$$\begin{aligned} \forall^\infty n : & \mathbb{P}\{f(x_n) = c, f_{n-1}(x_n, \tau'_n) \geq c - \delta\} \\ &= \mathbb{P}\{f(x_n) = c\} - \mathbb{P}\{f(x_n) = c, f_{n-1}(x_n, \tau'_n) < c - \delta\} \\ &> F(c) - F(c-) - \delta \geq F(c) - \mathbf{M}(\epsilon) - \delta. \end{aligned}$$

Now we are ready to analyze the composition of our five classes. Among the Class I extended examples at most

$$\delta n \quad (3.47)$$

will be above the threshold from some n on almost surely (by Lemma 3.11 and the Borel–Cantelli lemma). None of the Class II extended examples will be above the threshold, by definition. The fraction of Class III extended examples among the first n extended examples will tend to

$$F(c + 2\delta) - F(c) + F(c-) - F(c - 2\delta) \quad (3.48)$$

as $n \rightarrow \infty$ almost surely.

To estimate the number N_n^{IV} of Class IV extended examples among the first n extended examples, we use McDiarmid’s theorem. If one extended example is replaced by another, N_n^{IV} will change by at most $2K_n + 1$ (since this extended example can affect $f_n^{\neq}(x_i, \tau_i)$ for at most $2K_n$ other extended examples (x_i, τ_i, y_i)). Therefore,

$$\mathbb{P}\left\{\left|\frac{1}{n}N_n^{\text{IV}} - \frac{1}{n}\mathbb{E} N_n^{\text{IV}}\right| \geq \delta\right\} \leq 2e^{-2\delta^2 n/(2K_n+1)^2};$$

the assumption $K_n = o(\sqrt{n/\ln n})$ and the Borel–Cantelli lemma imply that

$$\left| \frac{1}{n} N_n^{\text{IV}} - \frac{1}{n} \mathbb{E} N_n^{\text{IV}} \right| < \delta$$

from some n on almost surely. Since

$$\frac{1}{n} \mathbb{E} N_n^{\text{IV}} = \mathbb{P} \{ f(x_n) = c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n) \} \geq F(c) - \mathbf{M}(\epsilon) - \delta$$

(see (3.45)), we have

$$N_n^{\text{IV}} > (F(c) - \mathbf{M}(\epsilon) - 2\delta)n \quad (3.49)$$

from some n on almost surely. Of course, all these examples are above the threshold.

Now we estimate the number $N_n^{\text{IV,str}}$ of n -strange extended examples of Class IV. Again McDiarmid's theorem implies that

$$\left| \frac{1}{n} N_n^{\text{IV,str}} - \frac{1}{n} \mathbb{E} N_n^{\text{IV,str}} \right| < \delta$$

from some n on almost surely. Now, from some n on,

$$\begin{aligned} \frac{1}{n} \mathbb{E} N_n^{\text{IV,str}} &= \mathbb{P} \{ f(x_n) = c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n), \hat{y}_n(x_n, \tau'_n) \neq y_n \} \\ &= \mathbb{E} ((1 - Q(\hat{y}_n(x_n, \tau'_n) | x_n)) \mathbb{I}_{\{f(x_n)=c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n)\}}) \\ &\leq e^{-\delta^*(n-1)} + \delta + (1 - c + 2\delta) \\ &\quad \times \mathbb{P}\{f(x_n) = c, (f_{n-1}(x_n, \tau'_n), 1 - \tau''_n) > (c'_n, c''_n)\} \\ &= e^{-\delta^*(n-1)} + \delta + (1 - c + 2\delta)(F(c) - \mathbf{M}(\epsilon) - \delta) \end{aligned} \quad (3.50)$$

$$\leq (F(c) - \mathbf{M}(\epsilon))(1 - c) + 4\delta \quad (3.51)$$

in the case $F(c) > \mathbf{M}(\epsilon)$; the first inequality in this chain follows from Lemma 3.10: indeed, this lemma implies that, unless an event of the small probability $e^{-\delta^*(n-1)} + \delta$ happens,

$$\begin{aligned} Q(\hat{y}_n(x_n, \tau'_n) | x_n) &\geq Q_{n-1}(\hat{y}_n(x_n, \tau'_n) | x_n, \tau'_n) - \delta \\ &= f_{n-1}(x_n, \tau'_n) - \delta \geq f(x_n) - 2\delta. \end{aligned} \quad (3.52)$$

If $F(c) = \mathbf{M}(\epsilon)$, the lines (3.50) and (3.51) of that chain have to be changed to

$$\begin{aligned} &\leq e^{-\delta^*(n-1)} + \delta + (1 - c + 2\delta) \mathbb{P}\{f(x_n) = c, f_{n-1}(x_n, \tau'_n) \geq c + \delta\} \\ &\leq e^{-\delta^*(n-1)} + \delta + (1 - c + 2\delta) (e^{-\delta^*(n-1)} + \delta) < 4\delta \end{aligned}$$

(where the obvious modification of Lemma 3.10 with all " $> \delta$ " changed to " $\geq \delta$ " is used), but the inequality between the extreme terms of the chain still

holds. Therefore, the number of n -strange Class IV extended examples does not exceed

$$((F(c) - M(\epsilon))(1 - c) + 5\delta)n \quad (3.53)$$

from some n on almost surely.

By the Borel strong law of large numbers, the fraction of Class V extended examples among the first n extended examples will tend to

$$1 - F(c + 2\delta) \quad (3.54)$$

as $n \rightarrow \infty$ almost surely. By Lemma 3.11, the Borel–Cantelli lemma, and (3.46), almost surely from some n on at least

$$(1 - F(c + 2\delta) - 2\delta)n \quad (3.55)$$

extended examples in Class V will be above the threshold.

Finally, we estimate the number $N_n^{V,\text{str}}$ of n -strange extended examples of Class V among the first n extended examples. By McDiarmid’s theorem,

$$\left| \frac{1}{n} N_n^{V,\text{str}} - \frac{1}{n} \mathbb{E} N_n^{V,\text{str}} \right| < \delta$$

from some n on almost surely. Now

$$\begin{aligned} \frac{1}{n} \mathbb{E} N_n^{V,\text{str}} &= \mathbb{P} \{f(x_n) > c + 2\delta, \hat{y}_n(x_n, \tau'_n) \neq y_n\} \\ &= \mathbb{E} ((1 - Q(\hat{y}_n(x_n, \tau'_n) | x_n)) \mathbb{I}_{f(x_n) > c+2\delta}) \\ &\leq e^{-\delta^*(n-1)} + \delta + \mathbb{E} ((1 - f(x_n) + 2\delta) \mathbb{I}_{f(x_n) > c+2\delta}) \\ &\leq e^{-\delta^*(n-1)} + 3\delta + \mathbb{E} ((1 - f(x_n)) \mathbb{I}_{f(x_n) > c+2\delta}) \\ &= e^{-\delta^*(n-1)} + 3\delta + \int_0^1 (F(\beta) - F(c + 2\delta))^+ d\beta \\ &< \int_0^1 (F(\beta) - F(c))^+ d\beta + 4\delta \end{aligned}$$

from some n on (the first inequality follows from Lemma 3.10, as in (3.52)). Therefore,

$$\frac{1}{n} N_n^{V,\text{str}} < \int_0^1 (F(\beta) - F(c))^+ d\beta + 5\delta \quad (3.56)$$

from some n on almost surely.

Summarizing, we can see that the total number of extended examples above the threshold among the first n extended examples will be at least

$$\begin{aligned} (F(c) - M(\epsilon) - 2\delta + 1 - F(c + 2\delta) - 2\delta)n \\ = (1 - M(\epsilon) + F(c) - F(c + 2\delta) - 4\delta)n \quad (3.57) \end{aligned}$$

(see (3.49) and (3.55)) from some n on almost surely. The number of n -strange extended examples among them will not exceed

$$\begin{aligned} & \left(\delta + F(c+2\delta) - F(c) + F(c-) - F(c-2\delta) + \delta \right. \\ & \quad \left. + (F(c) - \mathbf{M}(\epsilon))(1-c) + 5\delta + \int_0^1 (F(\beta) - F(c))^+ d\beta + 5\delta \right) n \\ & = \left(F(c+2\delta) - F(c) + F(c-) - F(c-2\delta) \right. \\ & \quad \left. + (F(c) - \mathbf{M}(\epsilon))(1-c) + \int_0^1 (F(\beta) - F(c))^+ d\beta + 12\delta \right) n \quad (3.58) \end{aligned}$$

(see (3.47), (3.48), (3.53), and (3.56)) from some n on almost surely. Combining equations (3.57) and (3.58), we can see that the number of n -strange extended examples among the $\lfloor (1 - \mathbf{M}(\epsilon))n \rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$ does not exceed

$$\begin{aligned} & \left(F(c+2\delta) - F(c) + F(c-) - F(c-2\delta) + (F(c) - \mathbf{M}(\epsilon))(1-c) \right. \\ & \quad \left. + \int_0^1 (F(\beta) - F(c))^+ d\beta + 12\delta \right) n + (F(c+2\delta) - F(c) + 4\delta) n \\ & = \left(2(F(c+2\delta) - F(c)) + (F(c-) - F(c-2\delta)) + (F(c) - \mathbf{M}(\epsilon))(1-c) \right. \\ & \quad \left. + \int_0^1 (F(\beta) - F(c))^+ d\beta + 16\delta \right) n \end{aligned}$$

from some n on almost surely. Since δ can be arbitrarily small, the coefficient in front of n in the last expression can be made arbitrarily close to

$$(F(c) - \mathbf{M}(\epsilon))(1-c) + \int_0^1 (F(\beta) - F(c))^+ d\beta = \int_0^1 (F(\beta) - \mathbf{M}(\epsilon))^+ d\beta = \epsilon,$$

which completes the proof. \square

Lemma 3.13. *Suppose (3.13) is satisfied. The fraction of n -strange extended examples among the first n extended examples (x_i, τ_i, y_i) approaches ϵ_0 asymptotically with probability one.*

Proof sketch. The lemma is not difficult to prove using McDiarmid's theorem and the fact that, by Lemma 3.11, $Q(\hat{y}_n(x_i, \tau'_i) | x_i)$ will typically differ little from $f(x_i)$. Notice, however, that the part that we really need (that the fraction of n -strange extended examples does not exceed $\epsilon_0 + o(1)$ as $n \rightarrow \infty$ with probability one) is just a special case of Lemma 3.12, corresponding to $\epsilon = \epsilon_0$. \square

Lemma 3.14. Suppose that (3.13) is satisfied and $\epsilon > \epsilon_0$. The fraction of n -conforming extended examples among the $\lfloor E(\epsilon)n \rfloor$ extended examples (x_i, τ_i, y_i) , $i = 1, \dots, n$, with the lowest $(f_n^\#(x_i, \tau'_i), 1 - \tau''_i)$ does not exceed $\epsilon - \epsilon_0 + o(1)$ as $n \rightarrow \infty$ with probability one.

Lemma 3.14 can be proved analogously to Lemma 3.12.

Lemma 3.15. Let $\mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \dots$ be a decreasing sequence of σ -algebras and ξ_1, ξ_2, \dots be a bounded adapted (in the sense that ξ_n is \mathcal{F}_n -measurable for all n) sequence of random variables such that

$$\limsup_{n \rightarrow \infty} \mathbb{E}(\xi_n | \mathcal{F}_{n+1}) \leq 0 \quad a.s.$$

Then

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_i \leq 0 \quad a.s.$$

Proof. Replacing, if necessary, ξ_n by $\xi_n - \mathbb{E}(\xi_n | \mathcal{F}_{n+1})$, we reduce our task to the following special case (a reverse Borel strong law of large numbers): if ξ_1, ξ_2, \dots is a bounded *reverse martingale difference*, in the sense of being adapted and satisfying $\forall n : \mathbb{E}(\xi_n | \mathcal{F}_{n+1}) = 0$, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_i = 0 \quad a.s. \tag{3.59}$$

Fix a bounded reverse martingale difference ξ_1, ξ_2, \dots ; our goal is to prove (3.59). By (the martingale version of) Hoeffding's inequality (see §A.7) applied to the martingale difference (ξ_i, \mathcal{F}_i) , $i = n, \dots, 1$,

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \xi_i \right| \geq \delta \right\} \leq 2e^{-\delta^2 n / (2C^2)}, \tag{3.60}$$

where C is an upper bound on $\sup_n |\xi_n|$. Combined with the Borel–Cantelli–Lévy lemma, (3.60) implies (3.59). \square

Now we can sketch the proof of Proposition 3.5. Define \mathcal{F}_n , $n = 1, 2, \dots$, to be the σ -algebra on $\tilde{\mathbf{Z}}^\infty$ generated by the bag of the first $n - 1$ extended examples (x_i, τ_i, y_i) , $i = 1, \dots, n - 1$, and the sequence of extended examples (x_i, τ_i, y_i) , $i = n, n + 1, \dots$ (starting from the n th extended example).

Suppose first that $\epsilon < \epsilon_0$. Consider the $\lfloor (1 - M(\epsilon - \delta))n \rfloor$ extended examples with the largest $(f_n^\#(x_i, \tau'_i), 1 - \tau''_i)$ among $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$, where $\delta \in (0, \epsilon)$ is a small constant. Let us show that each of these examples will be predicted with a non-multiple prediction from the other extended examples in the sequence $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$, from some n on. We will assume n large enough.

Let (x_k, τ_k, y_k) be the extended example with the $(\lfloor(\epsilon - \delta/2)n\rfloor + 1)$ th largest (in the sense of the lexicographic order) $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$ among all n -strange extended examples (x_i, τ_i, y_i) , $i = 1, \dots, n$. (Remember that all τ''_i are assumed to be different.) Let (x_j, τ_j, y_j) be one of the $\lfloor(1 - M(\epsilon - \delta))n\rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$ and let $y \in \mathbf{Y}$ be a label different from $\hat{y}_n(x_j, \tau'_j)$. It suffices to prove that

$$|\{i = 1, \dots, n : \alpha_i^y > \alpha_j^y\}| + \tau''_j |\{i = 1, \dots, n : \alpha_i^y = \alpha_j^y\}| \leq n\epsilon$$

(cf. (3.12) on p. 62), where all α^y are computed as α in (3.27) (p. 75) from the sequence $(x_1, \tau_1, y_1), \dots, (x_n, \tau_n, y_n)$ with y_j replaced by y . Since $\alpha_j^y = f_n^{\neq}(x_j, \tau'_j)$ and $\alpha_i^y \neq \alpha_i$ for at most $2K_n + 1$ values of i (indeed, changing y_j will affect at most $2K_n + 1$ nonconformity scores), it suffices to prove

$$|\{i : \alpha_i > f_n^{\neq}(x_j, \tau'_j)\}| + \tau''_j |\{i : \alpha_i = f_n^{\neq}(x_j, \tau'_j)\}| \leq n(\epsilon - \delta^*) , \quad (3.61)$$

where $\delta^* \ll \delta$ is a positive constant.

Since $(f_n^{\neq}(x_j, \tau'_j), 1 - \tau''_j) \geq (\alpha_k, 1 - \tau''_k)$ (indeed, by Lemma 3.12, there are less than $(\epsilon - \delta/2)n$ n -strange extended examples among the $\lfloor(1 - M(\epsilon - \delta))n\rfloor$ extended examples with the largest $(f_n^{\neq}(x_i, \tau'_i), 1 - \tau''_i)$), (3.61) will follow from

$$|\{i : \alpha_i > \alpha_k\}| + \tau''_k |\{i : \alpha_i = \alpha_k\}| \leq n(\epsilon - \delta^*) . \quad (3.62)$$

If $|\{i : \alpha_i = \alpha_k\}| \leq \frac{\delta}{3}n$, the left-hand side of (3.62) does not exceed

$$\left(\epsilon - \frac{\delta}{2}\right)n + \frac{\delta}{3}n < n(\epsilon - \delta^*) ,$$

so we can, and will, assume without loss of generality that

$$|\{i : \alpha_i = \alpha_k\}| > \frac{\delta}{3}n . \quad (3.63)$$

Since τ''_i for the extended examples satisfying $\alpha_i = \alpha_k$ are output according to the uniform distribution \mathbf{U} , the expected value of τ''_k is about

$$\frac{(\epsilon - \delta/2)n - |\{i : \alpha_i > \alpha_k\}|}{|\{i : \alpha_i = \alpha_k\}|} ,$$

and so by Hoeffding's inequality and the Borel–Cantelli lemma we will have (from some n on)

$$\tau''_k \leq \frac{(\epsilon - \delta/2)n - |\{i : \alpha_i > \alpha_k\}|}{|\{i : \alpha_i = \alpha_k\}|} + \delta^* \quad (3.64)$$

(remember (3.63)). Equation (3.62) will hold because its left-hand side can be transformed using (3.64) as

$$\begin{aligned} |\{i : \alpha_i > \alpha_k\}| + \tau''_k |\{i : \alpha_i = \alpha_k\}| &\leq (\epsilon - \delta/2)n + \delta^* |\{i : \alpha_i = \alpha_k\}| \\ &\leq (\epsilon - \delta/2 + \delta^*)n \leq (\epsilon - \delta^*)n. \end{aligned}$$

The assertion we have just proved means that, almost surely from some n on,

$$\mathbb{P}(\{\text{mult}_n = 0\} | \mathcal{F}_{n+1}) \geq \frac{\lfloor (1 - \mathbf{M}(\epsilon - \delta))n \rfloor}{n} \geq 1 - \mathbf{M}(\epsilon - \delta) - \frac{1}{n}.$$

Since δ can be arbitrarily small and \mathbf{M} is continuous (Lemma 3.7), this implies

$$\limsup_{n \rightarrow \infty} \mathbb{E}(\text{mult}_n | \mathcal{F}_{n+1}) \leq \mathbf{M}(\epsilon) \quad \text{a.s.}$$

By Lemma 3.15 this implies, in turn,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{mult}_i \leq \mathbf{M}(\epsilon) \quad \text{a.s.},$$

which coincides with (3.22) (p. 69).

If $\epsilon \geq \epsilon_0$, Lemma 3.13 implies that

$$\lim_{n \rightarrow \infty} \mathbb{E}(\text{mult}_n | \mathcal{F}_{n+1}) = 0 \quad \text{a.s.};$$

in combination with Lemma 3.15 this again implies (3.22).

Inequality (3.23) is treated in a similar way to (3.22). Lemmas 3.13 and 3.14 imply that

$$\liminf_{n \rightarrow \infty} \mathbb{E}(\text{emp}_n | \mathcal{F}_{n+1}) \geq \mathbf{E}(\epsilon) \quad \text{a.s.} \tag{3.65}$$

(this inequality is vacuously true when $\epsilon \leq \epsilon_0$). Another application of Lemma 3.15 gives

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{emp}_i \geq \mathbf{E}(\epsilon) \quad \text{a.s.},$$

i.e., (3.23).

Remark The derivation of Proposition 3.5 from Lemmas 3.12–3.15 would be very simple if we defined the nonconformity measure by, say,

$$A(B, (x, \sigma, y)) := \begin{cases} (-\hat{f}_B(x, \sigma), \sigma) & \text{if } y = \hat{y}_B(x, \sigma) \\ (\hat{f}_B(x, \sigma), \sigma) & \text{otherwise} \end{cases}$$

(with the lexicographic order on nonconformity scores) instead of (3.15) (in which case the second addend in the numerator of (3.12) would be just τ''_n almost surely). Our definition (3.15), however, is simpler and, most importantly, facilitates the proof of Proposition 3.2. Another simplification would be to use Lemma 3.12 (applied to $\epsilon := \epsilon - \mathbf{E}(\epsilon)$) instead of Lemma 3.14 in the derivation of (3.65); we preferred a more symmetric picture.

3.6 Bibliographical remarks

Examples of nonconformity measures

For a derivation of the dual problem (3.6) (p. 58), see Vapnik 1998. The objects x_i with $\alpha_i > 0$ are known as *support vectors*; this is the origin of the name “support vector machine”.

The idea of reducing binary classification to regression is an old one. In the case of simple prediction the procedure is as follows: encode the labels as real numbers, one negative and the other positive, apply a regression algorithm, and define \hat{y}_n to be the “positive” label if the value predicted for y_n by the regression algorithm is positive, to be the “negative” label if the value predicted for y_n by the regression algorithm is negative, and define \hat{y}_n arbitrarily if the value predicted for y_n by the regression algorithm is zero. Probably the earliest suggestion of this kind was Fisher’s *discriminant analysis* (Fisher 1973b, §49.2): if there are, say, l_1 males and l_2 females in the training set and $l_1 + l_2 = l$, encode males as l_2/l and encode females as $-l_1/l$ (so that the mean of the encodings over the training set is 0), and use the least squares algorithm as the regression algorithm.

The precursor of conformal predictor suggested in Gammerman et al. 1998 used the SVM method as the underlying algorithm. Later it was noticed (see Saunders et al. 1998) that the Lagrange method applied to ridge regression in analogy with SVM leads to α_i equivalent to the residuals, and this in turn lead to the realization that almost any machine learning algorithm can be adapted, often in more than one way, to obtain a nonconformity measure. However, the first genuine conformal predictor (then called “transductive confidence machine”) introduced in Vovk et al. 1999 and Saunders et al. 1999 still used the Lagrange multipliers α_i corresponding to constraints (3.4) (p. 57) as the nonconformity measure. The original conformal predictors for multilabel classification problems using binary SVMs were based on (3.7) with $\lambda = 1$, but it was quickly noticed that taking $\lambda < 1$ improves results dramatically.

There is a version of SVM for regression (see Vapnik 1998, Chap. 11), which can also be used for computing nonconformity scores.

We mentioned two methods of reducing multilabel classification problems to binary ones: “one-against-the-rest” and “one-against-one”. Another popular method, based on error-correcting coding, was proposed by Dietterich and Bakiri (1995).

Instead of reducing a multilabel classification problem to the binary case and then applying the SVM method, it is possible to use directly known multilabel generalizations of SVM. First such generalization was proposed by Blanz and Vapnik (Vapnik 1998, §10.10); later but independently it was found by Watkins and Weston (1999) and Jaakkola.

Universal predictor

The first step towards a universal predictor was done in Vovk 2002a, where it was shown that an optimal smoothed conformal predictor exists when the power distribution Q^∞ generating the examples is known. The full result was announced in Vovk 2003a.

Alternative protocols

Several papers (such as Rivest and Sloan 1988, Freund et al. 2004) extend the standard PAC framework by allowing the prediction algorithm to abstain from making a prediction at some trials. Our results show that for any significance level ϵ there exists a prediction algorithm that: (a) makes a wrong prediction with frequency at most ϵ ; (b) has an optimal frequency of abstentions among the prediction algorithms that satisfy property (a). The protocol of Rivest and Sloan (1988) and Freund et al. (2004) is in fact a restriction of our protocol, in which Predictor is only allowed to output a one-element set or the whole of \mathbf{Y} ; the latter is interpreted as abstention. (And in the situation where Err_n and Mult_n are of primary interest, as in this chapter, the difference between these two protocols is not very significant.) The universal predictor can be adapted to the restricted protocol by replacing a multiple prediction with \mathbf{Y} and replacing an empty prediction with a randomly chosen label. In this way we obtain a prediction algorithm in the restricted protocol which is asymptotically conservative and has an optimal frequency of abstentions, in the sense of (3.9) (p. 61), among the asymptotically conservative algorithms.

The methods of Freund et al. (2004) are directly applicable to conformal prediction; in particular, that paper defines a natural nonconformity measure (the “empirical log ratio”, taken with appropriate sign) in the situation where a hypothesis class is given.

Confidence and credibility

In the situation where Mult_n and Emp_n are the principal measures of predictive efficiency, it is very natural to summarize the range of possible prediction sets Γ^ϵ , $\epsilon \in (0, 1)$, by reporting the *confidence*

$$\sup\{1 - \epsilon : |\Gamma^\epsilon| \leq 1\}, \quad (3.66)$$

the *credibility*

$$\inf\{\epsilon : |\Gamma^\epsilon| = 0\},$$

and the *prediction* Γ^ϵ , where $1 - \epsilon$ is the confidence (in this case Γ^ϵ is never multiple for conformal predictors and usually contains exactly one label). Reporting the prediction, confidence, and credibility was suggested in Vovk et al. 1999 and Saunders et al. 1999; it is analogous to reporting the observed level of significance (Cox and Hinkley 1974, p. 66) in statistics.

Modifications of conformal predictors

So far we have emphasized desirable properties of conformal predictors: validity (Chap. 2), asymptotic efficiency (Chap. 3), and flexibility (ability to incorporate a wide range of machine-learning methods); we have also mentioned that the hedged predictions output by good conformal predictors are “conditional”, in the sense that they take full account of the object to be predicted. In this chapter we will discuss some limitations of conformal prediction and ways to overcome or alleviate these limitations.

The first problem, dealt with in §4.1, is the relative computational inefficiency of conformal predictors. In that section we construct “inductive conformal predictors” (ICP), whose computational efficiency is often much better; the price is some loss in predictive efficiency (which was called simply “efficiency” in the previous chapters). In §4.2 we introduce several new nonconformity measures, which are especially natural when used with ICP.

“Weak teachers”, which are allowed to provide the true label with a delay or not to provide it at all, are considered in §4.3. We introduce a formal notion of a “teaching schedule”, which is a fairly general protocol for disclosing labels of observed objects including several interesting special cases. The main result of that section is a characterization of teaching schedules under which the method of conformal prediction remains “asymptotically valid in probability”.

The protocol with a weak teacher is a relaxation of the pure on-line protocol in the direction of the off-line setting. After showing in §4.3 that conformal predictors retain some properties of validity in the mixed protocol, in §4.4 we discuss simple validity properties of off-line conformal predictors and inductive conformal predictors, and also briefly consider a mixed protocol for inductive conformal predictors.

The issue of conditionality is taken up in §4.5. The potentially serious problem with conformal predictors is that they are not automatically *conditionally valid*: e.g., in the USPS data set some digits (such as “5”) are more difficult to recognize correctly than other digits (such as “0”), and it is natural to expect that at the confidence level 95% the error rate will be significantly greater than 5% for the difficult digits; our usual, unconditional, notion of

validity only ensures that the average error rate over all digits will be close to 5%. The notion of Mondrian conformal predictor is introduced to address this concern.

4.1 Inductive conformal predictors

We start by looking more closely at the reasons for the relative computational inefficiency of conformal predictors for large data sets. For concreteness, we will discuss the conformal predictors determined by the simplest nonconformity measures (2.23) and (2.24) (p. 29), but the phenomenon is general. (The notion of ICP itself will be closer to conformal predictors determined by (2.24), in that ICPs never use the value of a prediction rule on examples from which the rule was found.)

As discussed in Chap. 1, one can usually assign a simple predictor to one of two types: “inductive” or “transductive”. For inductive predictors, $D_{\{z_1, \dots, z_n\}}$ can be computed, in some sense: e.g., $D_{\{z_1, \dots, z_n\}}$ may be described by a polynomial, and computing $D_{\{z_1, \dots, z_n\}}$ may mean computing the coefficients of the polynomial; as soon as $D_{\{z_1, \dots, z_n\}}$ is computed, computing $D_{\{z_1, \dots, z_n\}}(x)$ for a new object x takes very little time. For transductive predictors, relatively little can be done before seeing the new object x ; even allowing considerable time for pre-processing $\{z_1, \dots, z_n\}$, computing $D_{\{z_1, \dots, z_n\}}(x)$ will be a difficult task.

Notice that, even when D is an inductive algorithm, the confidence predictor based on the generic nonconformity measure (2.23) (and, even more so, on (2.24)) will still be computationally inefficient: for every new object x_n , computing $\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ will require constructing new prediction rules. Inductive conformal predictors will be defined in such a way that they can make significant computational savings when the underlying simple predictor D is inductive.

To define an ICP from a nonconformity measure (A_n) first fix a finite or infinite sequence of positive integer parameters m_1, m_2, \dots (called *update trials*); it is required that $m_1 < m_2 < \dots$. If the sequence m_1, m_2, \dots is finite, $(m_1, m_2, \dots) = (m_1, \dots, m_r)$, we set $m_i := \infty$ for $i > r$. The *ICP* determined by (A_n) and the sequence m_1, m_2, \dots of update trials is defined to be the confidence predictor Γ such that the prediction sets $\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ are computed as follows:

- if $n \leq m_1$, $\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ is found using a fixed conformal predictor;
- otherwise, find the k such that $m_k < n \leq m_{k+1}$ and set

$$\begin{aligned} \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) := \\ \left\{ y \in \mathbf{Y} : \frac{|\{j = m_k + 1, \dots, n : \alpha_j \geq \alpha_n\}|}{n - m_k} > \epsilon \right\}, \quad (4.1) \end{aligned}$$

where the nonconformity scores α_j are defined by

$$\begin{aligned}\alpha_j &:= A_{m_k+1}(\mathcal{I}(x_1, y_1), \dots, (x_{m_k}, y_{m_k}), (x_j, y_j)), \\ &\quad \text{for } j = m_k + 1, \dots, n - 1, \\ \alpha_n &:= A_{m_k+1}(\mathcal{I}(x_1, y_1), \dots, (x_{m_k}, y_{m_k}), (x_n, y)).\end{aligned}\quad (4.2)$$

Smoothed ICPs can be defined analogously to smoothed conformal predictors: instead of (4.1) we have

$$\begin{aligned}\Gamma^\epsilon(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) &:= \\ \left\{ y \in \mathbf{Y} : \frac{|\{j : \alpha_j > \alpha_n\}| + \tau_n |\{j : \alpha_j = \alpha_n\}|}{n - m_k} > \epsilon \right\},\end{aligned}\quad (4.3)$$

where $j = m_k + 1, \dots, n$ and $\tau_n \in [0, 1]$ are the random numbers. The following result (which is also a special case of Theorem 8.1 on p. 193) shows that Propositions 2.3 and 2.4 continue to hold in the case of ICPs and smoothed ICPs, respectively.

Proposition 4.1. *All ICPs are conservatively valid. All smoothed ICPs are exactly valid.*

The general scheme for defining nonconformity

For use with inductive confidence predictors, it is convenient to rewrite the definitions (2.23) and (2.24) more explicitly as

$$A(\mathcal{I}(x_1, y_1), \dots, (x_l, y_l), (x, y)) := \Delta(y, D_{\mathcal{I}(x_1, y_1), \dots, (x_l, y_l)}(x)) \quad (4.4)$$

and

$$A(\mathcal{I}(x_1, y_1), \dots, (x_l, y_l), (x, y)) := \Delta(y, D_{\mathcal{I}(x_1, y_1), \dots, (x_l, y_l)}(x)), \quad (4.5)$$

respectively. In the case where A is defined by (4.5), we can see that the ICP requires recomputing the prediction rule being used not at every trial but only at the update trials m_1, m_2, \dots ; the rate of growth of m_i determines the chosen balance between predictive and computational efficiency. The simplest nontrivial case, where there is only one update trial m_1 , is discussed in §4.4 below.

Let a and b be positive numbers such that either $a \geq 1$ and $b \geq 1$ or $a > 1$, and suppose that the prediction rule $D_{\mathcal{I}(z_1, \dots, z_n)}$ is computable in time $\Theta(n^a \log^b n)$ and the discrepancy measure Δ is computable in constant time. Then the conformal predictor determined by (4.5) spends time $\Theta(n^{a+1} \log^b n)$ on the computations needed for the first n trials. On the other hand, if the sequence m_i is infinite and grows exponentially fast, the ICP based on D , Δ , and (m_i) spends the same, to within a constant factor, time $\Theta(n^a \log^b n)$. (We have been assuming that the conformal predictor and ICP are given D as an

oracle and that the label space \mathbf{Y} is finite and fixed.) In the case where the sequence m_i is finite, the ICP's computation time becomes

$$\Theta(n \log n) \quad (4.6)$$

(e.g., use red-black trees for storing the nonconformity scores, as in the proof of Proposition 3.2 on p. 75, but augment them with information needed to find the rank of an element in time $O(\log n)$ – see Cormen et al. 2001, §14.1).

4.2 Further ways of computing nonconformity scores

All nonconformity measures described in the previous chapters can be used in inductive conformal prediction, and all nonconformity measures that will be introduced in this section can be used in conformal prediction. The nonconformity measures of this section are, however, especially convenient in the case of ICPs.

Suppose we are given a bag

$$\{z_1, \dots, z_l\} \in \mathbf{Z}^{(*)} \quad (4.7)$$

and an example $z \in \mathbf{Z}$, fixed for the rest of this section. The problem is to define the nonconformity score

$$A(\{z_1, \dots, z_l\}, z), \quad (4.8)$$

which we will usually abbreviate to $A(z)$. In the context of inductive conformal prediction, we are interested in $l = m_1, m_2, \dots$. Sometimes it will be more convenient to define the conformity score $B(z)$ instead. As usual, we write (x_i, y_i) for z_i and (x, y) for z when we need separate notations for the objects and labels.

We start from applying two nonconformal measures introduced earlier to ICP. The nonconformity measure used to define the deleted LSCM, $\alpha_i = |e_{(i)}|$ with the deleted residuals $e_{(i)}$ defined by (2.35) (p. 34), can be rewritten in our present context as

$$A(\{z_1, \dots, z_l\}, z) = |y - \hat{y}|, \quad (4.9)$$

where \hat{y} is the least squares prediction for y as computed from the training set $\{z_1, \dots, z_l\}$ and x . The nonconformity measure (2.36) used to define the studentized LSCM can be rewritten as

$$A(\{z_1, \dots, z_l\}, z) = \frac{|y - \hat{y}|}{\sqrt{1 + x'(X'X)^{-1}x}}, \quad (4.10)$$

where, in addition, X is the $l \times p$ matrix $(x_1, \dots, x_l)'$. This can be checked using the standard formula

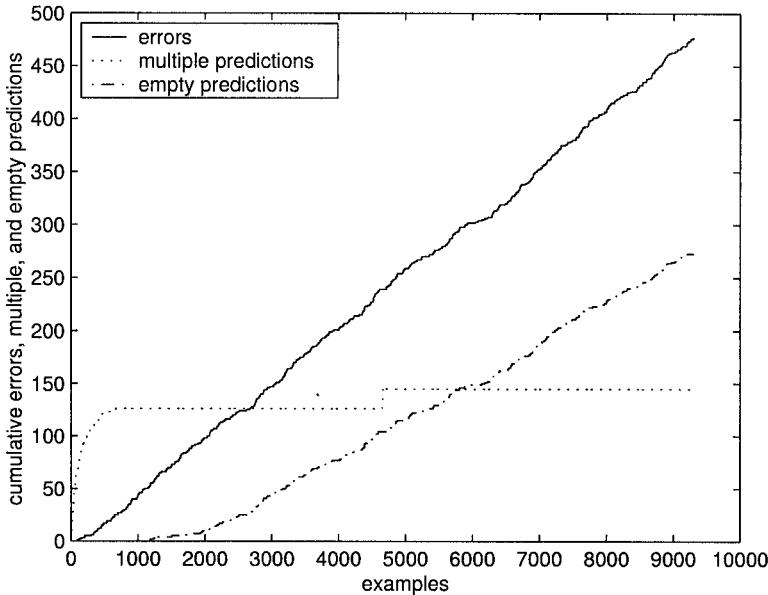


Fig. 4.1. On-line performance of the 1-nearest neighbor ICP with the update trial 4649 ($= 9298/2$) on the USPS data set for the confidence level 95%. In accordance with Proposition 4.1, starting from scratch at trial 4670 does not affect the error rate (solid line)

$$(K + uv')^{-1} = K^{-1} - \frac{K^{-1}uv'K^{-1}}{1 + v'K^{-1}u}, \quad (4.11)$$

where K is a square matrix and u and v are vectors.

The definition (3.1) (p. 54) of nonconformity measures based on the nearest neighbors classification is already given in the form (4.8) convenient for use with ICP. In the case of regression (p. 38), $A(x, y)$ is defined as $|y - \hat{y}|$, where \hat{y} is the k -NNR prediction for x computed from the bag (4.7).

The performance of the 1-nearest neighbor ICP on the USPS data set with update trial 4649 (the middle of the data set) is shown in Figs. 4.1 and 4.2. It can be seen from these figures (and is obvious anyway) that the ICP's performance (measured by the number of multiple predictions) deteriorates sharply after update trials m_i . (There is a hike of approximately $1/\epsilon$ in the number of multiple predictions, where ϵ is the significance level used.) Perhaps in practice there should be short spells of “learning” after each update trial, when the ICP is provided with fresh “training examples” and its predictions are not used or evaluated.

It is not clear how the way of computing nonconformity scores from SVM, as given in §3.1, could be used by ICP in a computationally efficient way. The easiest solution is perhaps to compute the SVM prediction rule based on the bag (4.7) and define $A(x, y)$ to be the distance (perhaps in a feature space)

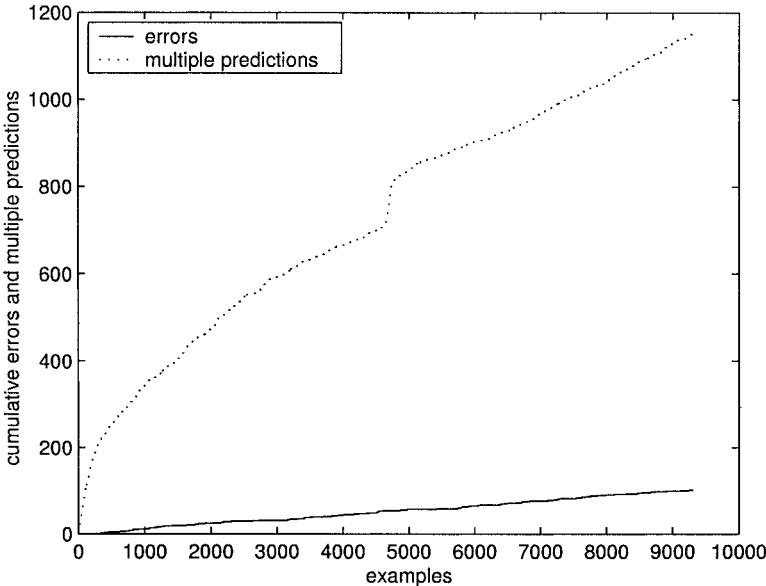


Fig. 4.2. On-line performance of the 1-nearest neighbor ICP with the update trial 4649 on the USPS data set for the confidence level 99%

between x and the optimal separating hyperplane (taken with the minus sign if the SVM prediction for x is different from y).

In the rest of this section we will describe new ways of detecting nonconformity which are especially natural in the case of inductive conformal prediction.

De-Bayesing

Suppose we have a Bayesian model (compatible with the randomness assumption) for the process of generating the label y given the object x . If we fully trust the model, we can use it for computing, e.g., predictive densities and prediction sets in the form of highest probability density regions (see, e.g., Bernardo and Smith 1994, §5.1). We are, however, interested in the case where the Bayesian model is plausible, but we do not really believe it. If it happens to be true, we would like our confidence predictor to be efficient. But we also want it to be always valid, even if the Bayesian model is wrong.

A natural definition of nonconformity measure (4.8) is as follows: find the posterior (after seeing the old examples z_1, \dots, z_l and the new object x) conditional distribution p for the label y given x , and define the conformity score for (x, y) as

$$B(\{z_1, \dots, z_l\}, (x, y)) := p\{y\} \quad (4.12)$$

in the case of classification (\mathbf{Y} is finite) and

$$B(\{z_1, \dots, z_l\}, (x, y)) := \min(p((-\infty, y]), p[y, \infty)) \quad (4.13)$$

in the case of regression ($\mathbf{Y} = \mathbb{R}$). In both cases, $B(x, y)$ is small when the label is strange under the Bayesian model, so the corresponding ICP is likely to be predictively efficient. The conditional probability distribution for the next example z given z_1, \dots, z_l can be computed before seeing x , which may lead to a computationally efficient ICP. And of course, the ICP will be valid automatically.

The ICP determined by one of these conformity measures may be said to be the result of “de-Bayesing” of the original Bayesian algorithm. More generally, we can also say that the RRCM algorithm of §2.3 is a de-Bayesed version of ridge regression (it is well known, and demonstrated in Chap. 10, that ridge regression is the Bayesian algorithm for a normal prior).

Bootstrap

The basic idea of bootstrap is to use resampling (sampling from the sample, obtaining what is called *bootstrap samples*) to get an idea of the variability of the value of interest (for details, see Efron and Tibshirani 1993, Davison and Hinkley 1997). Let us again consider the case of regression.

One way to implement this idea is as follows. Find the least squares weights $\hat{w} := (X'X)^{-1}X'Y$ from the training set (4.7), where X is the $l \times p$ matrix $(x_1, \dots, x_l)'$ of the objects in the training set (assuming the object space is $\mathbf{X} = \mathbb{R}^p$) and Y is the corresponding $l \times 1$ vector of the labels in the training set. Let \hat{G} be the uniform distribution on the *centered modified residuals* $r_i - \bar{r}$, where

$$r_i := \frac{y_i - \hat{y}_i}{\sqrt{1 - h_{ii}}}, \quad \hat{y}_i := \hat{w} \cdot x_i$$

(cf. (2.36) on p. 34), and

$$\bar{r} := \frac{1}{l} \sum_{i=1}^l r_i.$$

(That is, \hat{G} puts the same weight $1/l$ on each r_i .) Let ξ_r^* , $r = 1, 2, \dots$, be a sequence of independent random vectors in \mathbb{R}^l whose components are independent and distributed as \hat{G} . Obtain RM (where R should be large enough and $M = 1$ is acceptable) “prediction errors” $\delta_{r,m}^*$ in the following way:

FOR $r = 1, \dots, R$:

$$Y_r^* := X\hat{w} + \xi_r^*;$$

$$\hat{w}_r^* := (X'X)^{-1}X'Y_r^* \text{ (least squares estimate from } X \text{ and } Y_r^*);$$

FOR $m = 1, \dots, M$:

sample ϵ_m^* from \hat{G} ;

$$\delta_{r,m}^* := (\hat{w} \cdot x + \epsilon_m^*) - \hat{w}_r^* \cdot x$$

END FOR

END FOR.

From the prediction errors for the new object x we can compute the corresponding conformity score for the full example (x, y) as, e.g.,

$$B(x, y) := \min \left(\left| \{(r, m) : y - \hat{y} \leq \delta_{r,m}^*\} \right|, \left| \{(r, m) : y - \hat{y} \geq \delta_{r,m}^*\} \right| \right),$$

where $\hat{y} := \hat{w} \cdot x$.

Decision trees

A decision tree (for a detailed description see, e.g., Mitchell 1997, Chap. 3) is a way of classifying the objects into a finite number of classes. The classification is performed by testing the values of different attributes, but the details will not be important for us.

There are many methods of constructing a decision tree from a training set of examples. One of the most popular methods is Quinlan's (1993) C4.5, but again we do not need the precise details. We will assume that each class contains at least one object from the training set: if this is not the case, the decision tree can always be “pruned” to make sure this property holds.

After a decision tree is constructed from the training set, we can define a conformity score $B(x, y)$ of the new example (x, y) as the percentage of examples labeled as y among the training examples whose objects are classified in the same way as x by the decision tree.

Boosting

Boosting is, as its name suggests, a method for improving the performance of a given prediction algorithm, usually called the *weak learner*¹. As usual in the boosting literature, we will assume that the weak learner can be applied to the training set (4.7) in which each example z_i , $i = 1, \dots, l$, is taken with a nonnegative weight w_i , with the weights summing to 1. If a weak learner cannot process weighted examples, a bag of training examples of the same size l should be sampled from the probability distribution $D\{x_i\} := w_i$, and this bag is then used to train the weak learner. The output of the weak learner is a prediction rule $h : \mathbf{X} \rightarrow \mathbf{Y}$.

Let us assume, for simplicity, that $\mathbf{Y} = \{-1, 1\}$. One of the most popular boosting algorithms, AdaBoost.M1, works as follows:

start with the probability distribution $D_1\{i\} := 1/m$, $i = 1, \dots, m$;

FOR $t = 1, \dots, T$:

call the weak learner providing it with D_t ;

get back the prediction rule $h_t : \mathbf{X} \rightarrow \mathbf{Y}$;

compute the error $\epsilon_t := \sum_{i=1, \dots, l: h_t(x_i) \neq y_i} D_t\{i\}$;

$\alpha_t := \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$;

¹There is no connection between weak learners and our “weak teachers” considered in the next section.

```

update  $D_{t+1}\{i\} := D_t\{i\}e^{-\alpha_t y_i h_t(x_i)}/Z_t$ ,  $i = 1, \dots, l$ ,
      where  $Z_t$  is the normalizing constant
END FOR.

```

The normalizing constant Z_t is chosen to make D_{t+1} a probability distribution. The result of the boosting procedure is the function $f : \mathbf{X} \rightarrow \mathbb{R}$ defined by

$$f(x) := \frac{\sum_{t=1}^T \alpha_t h_t(x)}{\sum_{t=1}^T \alpha_t}.$$

The prediction for a new object x is computed as $\hat{y} := \text{sign } f(x)$. It can also be used to define the conformity score

$$B(x, y) := y f(x) \quad (4.14)$$

for an example (x, y) . This is a natural measure from the theoretical point of view (Schapire et al. 1998, Theorem 5) and gives reasonable empirical results on benchmark data sets (Proedrou 2003).

Another natural way to define the conformity score of an example (x, y) is to use a conformity measure for the weak learner. Suppose, e.g., that the weak learner is a method for constructing decision trees. Then we can define the conformity score of (x, y) as

$$B(x, y) := \sum_{t=1}^T \alpha_t B_t(x, y), \quad (4.15)$$

where $B_t(x, y)$ is the conformity score of (x, y) computed from h_t , as described in the previous subsection. No significant difference in the empirical performance of (4.14) and (4.15) was found in Proedrou 2003.

Neural networks

Let $|\mathbf{Y}| < \infty$ (neural networks are usually used for classification). When fed with an object $x \in \mathbf{X}$, a neural network outputs a set of numbers o_y , $y \in \mathbf{Y}$, such that o_y reflects the likelihood that y is x 's label. (See Mitchell 1997 for details.) Inductive conformal predictors determined by nonconformity scores

$$A(x, y) := \frac{\sum_{y' \in \mathbf{Y}: y' \neq y} o_{y'}}{o_y + \gamma}, \quad (4.16)$$

where $\gamma \geq 0$ is a suitably chosen parameter, have been shown to have a reasonable empirical performance (Papadopoulos 2004). Results change little if the \sum in (4.16) is replaced by \max .

Logistic regression

The logistic regression model is only applicable in the case $\mathbf{Y} = \{0, 1\}$ and $\mathbf{X} = \mathbb{R}^p$, for some p ; according to this model, the conditional probability that $y = 1$ given x for an example (x, y) is given by

$$\frac{e^{w \cdot x}}{1 + e^{w \cdot x}}$$

for some weight vector $w \in \mathbb{R}^p$. If \hat{w} is, e.g., the maximum likelihood estimate found from the bag (4.7), it is natural to use the nonconformity measure

$$A(x, y) := \begin{cases} 1 + e^{-\hat{w} \cdot x} & \text{if } y = 1 \\ 1 + e^{\hat{w} \cdot x} & \text{if } y = 0 \end{cases} \quad (4.17)$$

(i.e., $1/A(x, y)$ is the estimated probability of the observed y given the observed x for the current example).

Remembering that nonconformity scores can be subjected to a monotonic transformation without changing the prediction sets, we can simplify (4.17) to

$$A(x, y) := \begin{cases} -\hat{w} \cdot x & \text{if } y = 1 \\ \hat{w} \cdot x & \text{if } y = 0 . \end{cases}$$

4.3 Weak teachers

In the pure on-line setting, considered so far, we get an immediate feedback (the true label) for every example that we predict. This makes practical applications of this scenario questionable. Imagine, for example, a mail sorting center using an on-line prediction algorithm for zip code recognition; suppose the feedback about the “true” label comes from a human expert. If the feedback is given for every object x_i , there is no point in having the prediction algorithm: we can just as well use the label provided by the expert. It would help if the prediction algorithm could still work well, in particular be valid, if only every, say, tenth object were classified by a human expert. Alternatively, even if the prediction algorithm requires the knowledge of all labels, it might still be useful if the labels were allowed to be given not immediately but with a delay (in our mail sorting example, such a delay might make sure that we hear from local post offices about any mistakes made before giving a feedback to the algorithm). In this section we will see that asymptotic validity still holds in many cases where missing labels and delays are allowed.

In the pure on-line protocol we had validity in the strongest possible sense: at each significance level ϵ each smoothed conformal predictor made errors independently with probability ϵ . Now we will not have validity in this strongest sense, and so we will consider three natural asymptotic definitions, requiring only that $\text{Err}_n^\epsilon / n \rightarrow \epsilon$ in a certain sense: weak validity, strong validity, and validity in the sense of the law of the iterated logarithm. Finally, we will prove a simple result about asymptotic efficiency.

Imperfectly taught predictors

We are interested in the protocol where the predictor receives the true labels y_n only for a subset of trials n , and even for this subset, y_n may be given with a delay. This is formalized by a function $\mathcal{L} : N \rightarrow \mathbb{N}$ defined on an infinite set $N \subseteq \mathbb{N}$ and required to satisfy

$$\mathcal{L}(n) \leq n$$

for all $n \in N$ and

$$m \neq n \implies \mathcal{L}(m) \neq \mathcal{L}(n)$$

for all $m \in N$ and $n \in N$; a function satisfying these properties will be called a *teaching schedule*. A teaching schedule \mathcal{L} describes the way the data is disclosed to the predictor: at the end of trial $n \in N$ it is given the label $y_{\mathcal{L}(n)}$ for the object $x_{\mathcal{L}(n)}$. The elements of \mathcal{L} 's domain N in the increasing order will be denoted $n_i: N = \{n_1, n_2, \dots\}$ and $n_1 < n_2 < \dots$. We denote the total number of labels disclosed by the beginning of trial n to a predictor taught according to the teaching schedule \mathcal{L} by $s(n) := |\{i : i \in N, i < n\}|$.

Let Γ be a confidence predictor and \mathcal{L} be a teaching schedule. The \mathcal{L} -taught version $\Gamma^{\mathcal{L}}$ of Γ is

$$\begin{aligned} \Gamma^{\mathcal{L}, \epsilon}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \\ = \Gamma^\epsilon(x_{\mathcal{L}(n_1)}, y_{\mathcal{L}(n_1)}, \dots, x_{\mathcal{L}(n_{s(n)})}, y_{\mathcal{L}(n_{s(n)})}, x_n). \end{aligned}$$

Intuitively, at the end of trial n the predictor $\Gamma^{\mathcal{L}}$ learns the label $y_{\mathcal{L}(n)}$ if $n \in N$ and learns nothing otherwise. An \mathcal{L} -taught (smoothed) conformal predictor is a confidence predictor that can be represented as $\Gamma^{\mathcal{L}}$ for some (smoothed) conformal predictor Γ .

Let us now consider several examples of teaching schedules.

Ideal teacher. If $N = \mathbb{N}$ and $\mathcal{L}(n) = n$ for each $n \in N$, then $\Gamma^{\mathcal{L}} = \Gamma$.

Slow teacher with a fixed lag. If $N = \{l+1, l+2, \dots\}$ for some $l \in \mathbb{N}$ and $\mathcal{L}(n) = n - l$ for all $n \in N$, then $\Gamma^{\mathcal{L}}$ is a predictor which learns true labels with a delay of l .

Slow teacher. The previous example can be generalized as follows. Let $l(n) = n + \text{lag}(n)$ where $\text{lag} : \mathbb{N} \rightarrow \mathbb{N}$ is an increasing function. Define $N := l(\mathbb{N})$ and $\mathcal{L}(n) := l^{-1}(n)$, $n \in N$. Then $\Gamma^{\mathcal{L}}$ is a predictor which learns the true label for each object x_n with a delay of $\text{lag}(n)$.

Lazy teacher. If $N \neq \mathbb{N}$ and $\mathcal{L}(n) = n$, $n \in N$, then $\Gamma^{\mathcal{L}}$ is given the true labels immediately but not for every object.

All results of this section (as will be clear from the proofs, given in §4.6) use only the following properties of smoothed conformal predictors Γ :

- At each significance level ϵ , the errors $\text{err}_n^\epsilon(\Gamma)$, $n = 1, 2, \dots$, are independent Bernoulli random variables with parameter ϵ .

- The predictions do not depend on the order of the examples learnt so far:

$$\Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) = \Gamma^\epsilon(x_{\pi(1)}, y_{\pi(1)}, \dots, x_{\pi(n-1)}, y_{\pi(n-1)}, x_n)$$

for any permutation π of $\{1, \dots, n-1\}$. (Remember that we call confidence predictors satisfying this property *invariant*.)

Weak validity

In this subsection we state a necessary and sufficient condition for a teaching schedule to preserve “weak asymptotic validity” of conformal predictors. The condition turns out to be rather weak: feedback should be given at more than a logarithmic fraction of trials.

We start from the definitions, assuming, for simplicity, randomness rather than exchangeability. A randomized confidence predictor Γ is *asymptotically exact in probability* if, for all significance levels ϵ and all probability distributions Q on \mathbf{Z} ,

$$\frac{1}{n} \sum_{i=1}^n \text{Err}_n^\epsilon(\Gamma, Q^\infty) - \epsilon \rightarrow 0$$

in probability. Similarly, a confidence predictor Γ is *asymptotically conservative in probability* if, for all significance levels ϵ and all probability distributions Q on \mathbf{Z} ,

$$\left(\frac{1}{n} \sum_{i=1}^n \text{Err}_n^\epsilon(\Gamma, Q^\infty) - \epsilon \right)^+ \rightarrow 0$$

in probability.

Theorem 4.2. *Let \mathcal{L} be a teaching schedule with domain $N = \{n_1, n_2, \dots\}$, where n_1, n_2, \dots is a strictly increasing infinite sequence of positive integers.*

- *If $\lim_{k \rightarrow \infty} (n_k/n_{k-1}) = 1$, any \mathcal{L} -taught smoothed conformal predictor is asymptotically exact in probability.*
- *If $\lim_{k \rightarrow \infty} (n_k/n_{k-1}) = 1$ does not hold, there exists an \mathcal{L} -taught smoothed conformal predictor which is not asymptotically exact in probability.*

In words, this theorem asserts that an \mathcal{L} -taught smoothed conformal predictor is guaranteed to be asymptotically exact in probability if and only if the growth rate of n_k is sub-exponential.

Corollary 4.3. *If $\lim_{k \rightarrow \infty} (n_k/n_{k-1}) = 1$, any \mathcal{L} -taught conformal predictor is asymptotically conservative in probability.*

Strong validity

Theorem 4.4. Suppose the example space \mathbf{Z} is Borel. Let Γ be a smoothed conformal predictor and \mathcal{L} be a teaching schedule whose domain is $N = \{n_1, n_2, \dots\}$, where $n_1 < n_2 < \dots$. If

$$\sum_k \left(\frac{n_k}{n_{k-1}} - 1 \right)^2 < \infty \quad (4.18)$$

then $\Gamma^{\mathcal{L}}$ is asymptotically exact.

This theorem shows that $\Gamma^{\mathcal{L}}$ is asymptotically exact when n_k grows as $\exp(\sqrt{k}/\ln k)$; on the other hand, it does not guarantee that it is asymptotically exact if n_k grows as $\exp(\sqrt{k})$.

Corollary 4.5. Let Γ be a conformal predictor and \mathcal{L} be a teaching schedule with domain $N = \{n_1, n_2, \dots\}$, $n_1 < n_2 < \dots$. Under condition (4.18), $\Gamma^{\mathcal{L}}$ is an asymptotically conservative confidence predictor.

Iterated logarithm validity

The following result asserts, in particular, that when n_k are equally spaced a stronger version of asymptotic validity, in the spirit of the law of the iterated logarithm, holds.

Theorem 4.6. Suppose the domain $\{n_1, n_2, \dots\}$, $n_1 < n_2 < \dots$, of a teaching schedule satisfies $n_k = O(k)$. Each \mathcal{L} -taught smoothed conformal predictor $\Gamma^{\mathcal{L}}$ satisfies

$$\left| \frac{\text{Err}_n^\epsilon(\Gamma^{\mathcal{L}}, Q^\infty)}{n} - \epsilon \right| = O\left(\sqrt{\frac{\ln \ln n}{n}}\right) \quad \text{a.s.}$$

and each \mathcal{L} -taught conformal predictor $\Gamma^{\mathcal{L}}$ satisfies

$$\left(\frac{\text{Err}_n^\epsilon(\Gamma^{\mathcal{L}}, Q^\infty)}{n} - \epsilon \right)^+ = O\left(\sqrt{\frac{\ln \ln n}{n}}\right) \quad \text{a.s. ,}$$

for each $Q \in \mathbf{P}(\mathbf{Z})$ at each significance level ϵ .

Efficiency

We will only consider the case of classification, taking the number of multiple predictions as the primary measure of inefficiency.

If Γ is a confidence predictor and ϵ a significance level, we set

$$U^\epsilon(\Gamma) = \left[\liminf_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma)}{n}, \limsup_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma)}{n} \right] .$$

The intervals $U^\epsilon(\Gamma)$ characterize the asymptotical efficiency of Γ ; of course, these are random intervals, since they depend on the actual examples output by Reality. It turns out, however, that in the most important case (covering conformal predictors and \mathcal{L} -taught conformal predictors, smoothed and deterministic) these intervals are close to being deterministic under the assumption of randomness.

Lemma 4.7. *For each invariant confidence predictor Γ (randomized or deterministic), significance level ϵ , and probability distribution Q on \mathbf{Z} there exists an interval $[a, b] \subseteq (0, 1)$ such that*

$$U^\epsilon(\Gamma) = [a, b] \quad a.s. ,$$

provided the examples and random numbers (if applicable) are generated from Q^∞ and \mathbf{U}^∞ independently.

Proof. The statement of this lemma is an immediate consequence of the Hewitt–Savage zero-one law (see, e.g., Shiryaev 1996, Theorem IV.1.3). \square

We will use the notation $U^\epsilon(\Gamma, Q)$ for the interval whose existence is asserted in the lemma; it characterizes the asymptotical efficiency of Γ at significance level ϵ with examples distributed according to Q .

Theorem 4.8. *Let Γ be a (smoothed) conformal predictor and \mathcal{L} be a teaching schedule defined on $N = \{n_1, n_2, \dots\}$, where $n_1 < n_2 < \dots$ is an increasing sequence. If, for some $c \in \mathbb{N}$, $n_{k+1} - n_k = c$ from some k on, then $U^\epsilon(\Gamma^{\mathcal{L}}, Q) = U^\epsilon(\Gamma, Q)$ for all significance levels $\epsilon \in (0, 1)$ and all probability distributions Q on \mathbf{Z} .*

Theorems 4.4 and 4.8 can be illustrated with the following simple example. Suppose only every m th label is revealed to a conformal predictor, and even this is done with a delay of l , where m and l are positive integer constants. Then (smoothed) conformal predictors will remain asymptotically valid, and their asymptotic rate of multiple predictions will not deteriorate.

4.4 Off-line conformal predictors and semi-off-line ICPs

As we discuss in this section, conformal predictors and ICPs can be applied in the pure off-line mode, but we will then only have a weakened guarantee of validity. The notion of ICP, however, has a natural “semi-off-line” version, which is exactly valid. This section’s discussion is independent of the previous section’s results (except for a short remark at the end), but it will be clear that they can be fruitfully combined.

Suppose we are given a training set z_1, \dots, z_l of examples $z_i = (x_i, y_i)$ and the problem is to predict the labels y_i , $i = l + 1, \dots, l + k$, of the *working examples* z_{l+1}, \dots, z_{l+k} . The *off-line conformal predictor* outputs the prediction sets

$$\begin{aligned}\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) := \\ \left\{ y \in \mathbf{Y} : \frac{|\{j = 1, 2, \dots, l, i : \alpha_j \geq \alpha_i\}|}{l+1} > \epsilon \right\} \quad (4.19)\end{aligned}$$

for each working example z_i , $i = l+1, \dots, l+k$, where the nonconformity scores are computed from a nonconformity measure A :

$$\begin{aligned}\alpha_j := A_{l+1}(\lfloor z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_l, (x_i, y) \rfloor, z_j), \quad j = 1, \dots, l, \\ \alpha_i := A_{l+1}(\lfloor z_1, \dots, z_l \rfloor, (x_i, y)) \quad (4.20)\end{aligned}$$

(cf. (2.18) and (2.19) on p. 26).

In a similar way we can define *off-line ICPs*. For concreteness, we restrict ourselves to the nonconformity measures (4.5) (p. 99). The training set is first split into two parts: the *proper training set*

$$(x_1, y_1, \dots, x_m, y_m) \quad (4.21)$$

of size $m < l$ and the *calibration set*

$$(x_{m+1}, y_{m+1}, \dots, x_l, y_l) \quad (4.22)$$

of size $l-m$. For every working object x_i , $i = l+1, \dots, l+k$, compute the prediction sets

$$\begin{aligned}\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) := \\ \left\{ y \in \mathbf{Y} : \frac{|\{j = m+1, \dots, l, i : \alpha_j \geq \alpha_i\}|}{l-m+1} > \epsilon \right\}, \quad (4.23)\end{aligned}$$

where the nonconformity scores are defined by

$$\begin{aligned}\alpha_j := \Delta(y_j, D_{\lfloor (x_1, y_1), \dots, (x_m, y_m) \rfloor}(x_j)), \quad j = m+1, \dots, l, \\ \alpha_i := \Delta(y, D_{\lfloor (x_1, y_1), \dots, (x_m, y_m) \rfloor}(x_i)). \quad (4.24)\end{aligned}$$

(Cf. (4.1)–(4.2), p. 99, and (4.5).)

For both conformal predictors and ICPs, it is true that

$$Q^\infty \{(x_1, y_1, x_2, y_2, \dots) : y_i \notin \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i)\} \leq \epsilon \quad (4.25)$$

for every $i = l+1, \dots, l+k$, provided all examples are drawn independently from the distribution Q , but the events in (4.25) are not independent and

$$\frac{|\{i = l+1, \dots, l+k : y_i \notin \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i)\}|}{k} \quad (4.26)$$

can be significantly above ϵ even when k is very large. (Cf. the description of the “inductivist objection” in §10.2.)

To ensure validity of the off-line ICP, we can modify the application of the ICP constructed from the training set to the working set: after processing

each working example (x_i, y_i) , $i = l + 1, \dots, l + k$, the corresponding nonconformity score α_i should be added to the pool of nonconformity scores used in generating the prediction sets for the following working examples. Formally, redefine

$$\begin{aligned} \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) := \\ \left\{ y \in \mathbf{Y} : \frac{|\{j = m + 1, \dots, i : \alpha_j \geq \alpha_i\}|}{i - m} > \epsilon \right\}, \end{aligned} \quad (4.27)$$

where the nonconformity scores are defined by

$$\begin{aligned} \alpha_j := \Delta(y_j, D_{l(x_1, y_1), \dots, (x_m, y_m)}(x_j)), \quad j = m + 1, \dots, i - 1, \\ \alpha_i := \Delta(y, D_{l(x_1, y_1), \dots, (x_m, y_m)}(x_i)). \end{aligned} \quad (4.28)$$

Proposition 4.1 (p. 99) says that this modification is conservatively valid, and so (4.26) will not exceed ϵ , up to statistical fluctuations.

Notice that in the case $k \ll (l - m)$ the *semi-off-line ICP* (4.27) differs so little from the off-line ICP that the latter can be expected to be “nearly conservative”.

Let us give a formal definition. A confidence predictor Γ is (δ_n) -conservative, where $\delta_1, \delta_2, \dots$ is a sequence of nonnegative numbers, if for any exchangeable probability distribution P on \mathbb{Z}^∞ there exists a probability space with two families

$$(\xi_n^{(\epsilon)} : \epsilon \in (0, 1), n = 1, 2, \dots), \quad (\eta_n^{(\epsilon)} : \epsilon \in (0, 1), n = 1, 2, \dots)$$

of $\{0, 1\}$ -valued random variables such that:

- for a fixed ϵ , $\xi_1^{(\epsilon)}, \xi_2^{(\epsilon)}, \dots$ is a sequence of independent Bernoulli random variables with parameter ϵ ;
- for all n and ϵ , $\eta_n^{(\epsilon-\delta_n)} \leq \xi_n^{(\epsilon)}$;
- the joint distribution of $\text{err}_n^\epsilon(\Gamma, P)$, $\epsilon \in (0, 1)$, $n = 1, 2, \dots$, coincides with the joint distribution of $\eta_n^{(\epsilon)}$, $\epsilon \in (0, 1)$, $n = 1, 2, \dots$.

The definition of conservative validity is a special case corresponding to $\delta_n = 0$, $n = 1, 2, \dots$; we are now interested in the case where δ_n are small (at least for a range of n) positive numbers.

Proposition 4.9. *The confidence predictor*

$$\tilde{\Gamma}^\epsilon(x_1, y_1, \dots, x_{i-1}, y_{i-1}, x_i) := \begin{cases} \Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i) & \text{if } i > l \\ \mathbf{Y} & \text{otherwise,} \end{cases}$$

where $\Gamma^\epsilon(x_1, y_1, \dots, x_l, y_l, x_i)$ is defined by (4.23), is (δ_i) -conservative, where

$$\delta_i := \begin{cases} \frac{i-l}{l-m} & \text{if } i > l \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Let Γ^\dagger be the smoothed semi-off-line ICP corresponding to Γ , and let $i > l$. Define ξ_i and η_i by the requirements that $\xi_i^{(\epsilon)} = 1$ if and only if Γ^\dagger makes a mistake when fed with $x_1, y_1, \dots, x_i, y_i$, and $\eta_i^{(\epsilon)} = 1$ if and only if Γ makes a mistake when fed with $x_1, y_1, \dots, x_l, y_l, x_i, y_i$, at the significance level ϵ . To ensure $\eta_i^{(\epsilon-\delta_i)} \leq \xi_i^{(\epsilon)}$, it is sufficient to require

$$(1 - \epsilon)(i - m) \leq (1 - \epsilon + \delta_i)(l - m + 1),$$

i.e.,

$$\delta_i \geq (1 - \epsilon) \frac{i - l - 1}{l - m + 1}. \quad \square$$

This proof shows that the fraction of errors made by the off-line ICP at a significance level ϵ on the working set does not exceed $\epsilon + k/(l - m)$, up to statistical fluctuations.

ICPs applied in both off-line and semi-off-line modes are computationally quite efficient. Let us see what the computation time will be if standard algorithms for standard computation tasks are used. In the case of simple predictions, the application of the inductive algorithm D found from the training set of size l to the working set of size k requires time

$$\Theta(T_{\text{train}} + kT_{\text{appl}}),$$

where T_{train} is the time required for computing the prediction rule $D_{\{z_1, \dots, z_l\}}$ and T_{appl} is the time needed to apply this prediction rule to a new object. The off-line ICP (see (4.23) and (4.24)) requires time

$$\Theta\left(T_{\text{train}}^\dagger + (l - m + k)T_{\text{appl}}^\dagger + (l - m)\log(l - m) + k\log(l - m)\right),$$

where T_{train}^\dagger is the time required for computing the prediction rule $D_{\{z_1, \dots, z_m\}}$ and T_{appl}^\dagger is the time needed to apply this prediction rule to a new object (we assume that computing Δ is fast); we allow time $(l - m)\log(l - m)$ for sorting the nonconformity scores obtained from the calibration set (Cormen et al. 2001, Part II) and time $\log(l - m)$ for finding the rank of a working nonconformity score in the set of the calibration nonconformity scores (Cormen et al. 2001, Chaps. 12 and 13). In the case of semi-off-line ICP ((4.27), (4.28)), the required time increases only slightly (for moderately large k) to

$$\Theta\left(T_{\text{train}}^\dagger + (l - m + k)T_{\text{appl}}^\dagger + (l - m)\log(l - m) + k\log(l - m + k)\right).$$

As $k \rightarrow \infty$, we have the same asymptotic computation time, $\Theta(k \log k)$, as in §4.1 (cf. (4.6) on p. 100). If, however, our goal is only asymptotic conservativeness as $k \rightarrow \infty$, by Theorem 4.4 we can keep only a fraction of $(\ln k)^3$ of the nonconformity scores α_i , $l < i \leq l + k$, and so the asymptotic computation time will become $\Theta(k \log \log k)$.

4.5 Mondrian conformal predictors

Our starting point in this section is a natural division of examples into several categories: e.g., different categories can correspond to different labels, or kinds of objects, or just be determined by the ordinal number of the example. As we have already discussed, conformal predictors do not guarantee validity within categories: the fraction of errors can be much larger than the nominal significance level for some categories, if this is compensated by a smaller fraction of errors for other categories. This stronger kind of validity, validity within categories, is the main property of Mondrian conformal predictors (MCPs), constructed in this section. As usual, we will demonstrate validity, in this stronger sense, under the exchangeability assumption; this assumption, however, will be relaxed in Chap. 8.

Validity within categories (or *conditional validity*, as we will say) is especially relevant in the situation of *asymmetric classification*, where errors for different categories of examples have different consequences; in this case we cannot allow low error rates for some categories to compensate excessive error rates for other categories. Because of our interest in asymmetric classification, we will mainly use the language of conformal transducers in our exposition. The standard translation into the language of conformal predictors is straightforward (cf. §2.5), but in the case of asymmetric classification one might prefer to add flexibility to this translation: instead of comparing all p-values with the same threshold ϵ we might take different ϵ s for different categories.

At the end of this section we discuss several special cases of MCPs, including conformal predictors and ICPs.

Mondrian conformal transducers

We are given a division of the Cartesian product $\mathbb{N} \times \mathbb{Z}$ into *categories*: a measurable function

$$\kappa : \mathbb{N} \times \mathbb{Z} \rightarrow K$$

maps each pair (n, z) (z is an example and n will be, in our applications, the ordinal number of this example in the data sequence z_1, z_2, \dots) to its category; K is the measurable space (at most countable with the discrete σ -algebra) of all categories. It is required that the elements $\kappa^{-1}(k)$ of each category $k \in K$ form a rectangle $A \times B$, for some $A \subseteq \mathbb{N}$ and $B \subseteq \mathbb{Z}$. Such a function κ will be called a *Mondrian taxonomy*.

Given a Mondrian taxonomy κ , we first define “Mondrian nonconformity measures” and then Mondrian conformal transducers (MCTs).

A *Mondrian nonconformity measure* based on κ is a family of measurable functions $(A_n : n \in \mathbb{N})$ of the type

$$A_n : K^{n-1} \times (\mathbb{Z}^{(*)})^K \times K \times \mathbb{Z} \rightarrow \overline{\mathbb{R}} .$$

The *smoothed Mondrian conformal transducer (smoothed MCT)* determined by the Mondrian nonconformity measure A_n is the randomized confidence transducer producing the p-values

$$\begin{aligned} p_n &= f(x_1, \tau_1, y_1, \dots, x_n, \tau_n, y_n) \\ &:= \frac{|\{i : \kappa_i = \kappa_n \& \alpha_i > \alpha_n\}| + \tau_n |\{i : \kappa_i = \kappa_n \& \alpha_i = \alpha_n\}|}{|\{i : \kappa_i = \kappa_n\}|}, \end{aligned} \quad (4.29)$$

where i ranges over $\{1, \dots, n\}$, $\kappa_i := \kappa(i, z_i)$, $z_i := (x_i, y_i)$, and

$$\begin{aligned} \alpha_i &:= A_n(\kappa_1, \dots, \kappa_{n-1}, \\ &\quad (k \mapsto \{z_j : j \in \{1, \dots, i-1, i+1, \dots, n\} \& \kappa_j = k\}), \kappa_n, z_i) \end{aligned} \quad (4.30)$$

for $i = 1, \dots, n$ such that $\kappa_i = \kappa_n$. As usual, the definition of a *Mondrian conformal transducer (MCT)* is obtained by replacing (4.29) with

$$p_n = f(x_1, y_1, \dots, x_n, y_n) := \frac{|\{i : \kappa_i = \kappa_n \& \alpha_i \geq \alpha_n\}|}{|\{i : \kappa_i = \kappa_n\}|}.$$

In general, a (*smoothed*) *MCT* based on a Mondrian taxonomy κ is the (*smoothed*) *MCT* determined by some Mondrian nonconformity measure based on κ .

We say that a randomized confidence transducer f is *category-wise exact w.r. to a Mondrian taxonomy κ* if, for all n , the conditional probability distribution of p_n given $\kappa(1, z_1), p_1, \dots, \kappa(n-1, z_{n-1}), p_{n-1}, \kappa(n, z_n)$ is uniform on $[0, 1]$, where z_1, z_2, \dots are examples generated from an exchangeable distribution on \mathbf{Z}^∞ and p_1, p_2, \dots are the p-values output by f .

Proposition 4.10. *Any smoothed MCT based on a Mondrian taxonomy κ is category-wise exact w.r. to κ .*

This proposition generalizes Proposition 4.1 but is a special case of Theorem 8.2 (the finitary version of Theorem 8.1) on p. 193. It implies the category-wise property of conservative validity for MCT, whose p-values are always bounded above by the p-values from the corresponding smoothed MCT.

Using Mondrian conformal transducers for prediction

An example of asymmetric classification is distinguishing between useful messages and spam in the problem of e-mail filtering: classifying a useful message as spam is a more serious error than vice versa. In this case we might want to have different significance levels ϵ_k for different categories k .

Let f be a (*smoothed*) MCT. Given a set of significance levels ϵ_k , $k \in K$, we can define the prediction set for the label y_n of a new object x_n given old examples z_1, \dots, z_{n-1} as

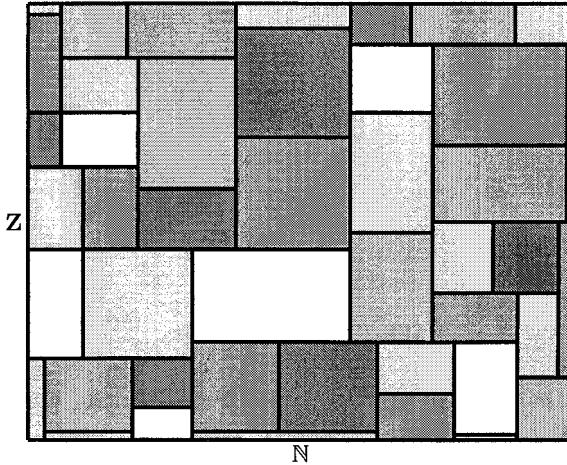


Fig. 4.3. A random Mondrian taxonomy (after Piet Mondrian, 1918)

$$\begin{aligned} \Gamma^{(\epsilon_k : k \in K)}(z_1, \dots, z_{n-1}, x_n) \\ := \{y \in \mathbf{Y} : f(z_1, \dots, z_{n-1}, (x_n, y)) > \epsilon_{\kappa(n, (x_n, y))}\} . \end{aligned}$$

Proposition 4.10 now implies that the long-run frequency of errors made by this predictor (*Mondrian conformal predictor*, or *MCP*) on examples of category k does not exceed (approaches, in the case of smoothed transducer) ϵ_k , for each k .

As in the case of conformal prediction, in applications it is usually not wise to fix thresholds ϵ_k , $k \in K$, in advance. One possibility would be to suitably choose three sets of significance levels (ϵ_k^1) , (ϵ_k^2) , and (ϵ_k^3) such that $\epsilon_k^1 \leq \epsilon_k^2 \leq \epsilon_k^3$ for all $k \in K$, and say that $\Gamma^{\epsilon^1}(z_1, \dots, z_{n-1}, x_n)$ is a highly confident prediction, $\Gamma^{\epsilon^2}(z_1, \dots, z_{n-1}, x_n)$ is a confident prediction, and $\Gamma^{\epsilon^3}(z_1, \dots, z_{n-1}, x_n)$ is a casual prediction.

Generality of Mondrian taxonomies

We will next consider several classes of MCTs, involving different taxonomies. In this subsection we consider a natural partial order on the taxonomies, which will clarify the relation between different special cases. (We will use the expression “more general than” for this partial order; it might seem strange here but will be explained in §8.4.) There are many ways to split the rectangle $\mathbb{N} \times \mathbb{Z}$ into smaller rectangles (cf. Fig. 4.3), and it is clearly desirable to impose some order.

We say that a Mondrian taxonomy κ_1 is *more general* than another Mondrian taxonomy κ_2 if, for all pairs (n', z') and (n'', z'') ,

$$\kappa_1(n', z') = \kappa_1(n'', z'') \implies \kappa_2(n', z') = \kappa_2(n'', z'') .$$

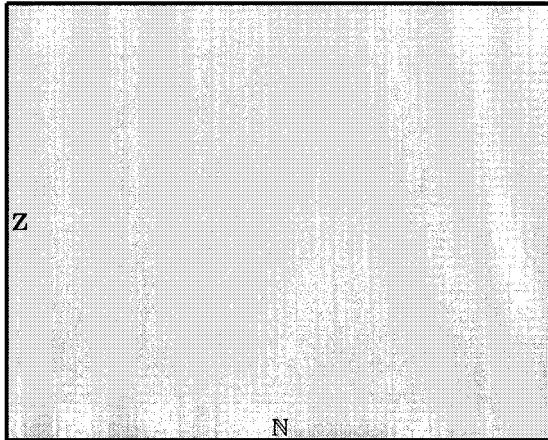


Fig. 4.4. Mondrian taxonomy corresponding to conformal transducers

We will say that κ_1 and κ_2 are *equivalent* if each of them is more general than the other, and we will sometimes identify equivalent Mondrian taxonomies. Identifying equivalent Mondrian taxonomies means that we are only interested in the equivalence relation a given Mondrian taxonomy κ induces ((n', z') and (n'', z'') are κ -*equivalent* if $\kappa(n', z') = \kappa(n'', z'')$) and not in the chosen labels $\kappa(n, z)$ for the equivalence classes.

Since we are only interested in taxonomies with at most countable number of categories, the following proposition immediately follows from the standard properties of conditional expectations (see property 2 on p. 280).

Proposition 4.11. *Let a taxonomy κ_1 be more general than a taxonomy κ_2 . If a randomized confidence transducer is category-wise exact w.r. to κ_1 , it is category-wise exact w.r. to κ_2 .*

Conformal transducers

Conformal transducers are MCTs based on the least general (i.e., constant, see Fig. 4.4) Mondrian taxonomy. Proposition 2.4 (p. 27), asserting that smoothed conformal predictors are exact, is a special case of Proposition 4.10.

In the rest of this section we will describe several experimental results for the USPS data set (randomly permuted), using the 1-nearest neighbor ratio (3.1) (p. 54) as the nonconformity measure. We start from results demonstrating the lack of conditional validity for conformal predictors. The USPS data set is reasonably balanced in the proportion of examples labeled by different digits; for less well-balanced data sets the lack conditional validity of non-Mondrian conformal predictors is often even more pronounced.

Figure 4.5 (plotting Err_n , ϵn , Mult_n , and Emp_n against n for the confidence level 95%; the plots for Err_n , Mult_n , and Emp_n are almost indistinguishable from the analogous plots for the deterministic conformal predictor)

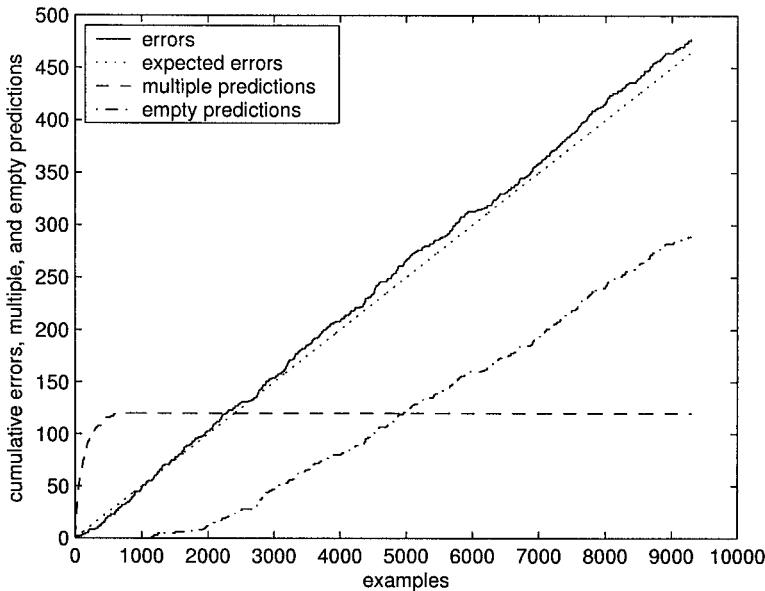


Fig. 4.5. The performance of the smoothed conformal predictor on the USPS data set at the 95% confidence level

shows that the smoothed conformal predictor is valid “on average” on the USPS data set.

Figure 4.6 gives similar plots, but only taking into account the predictions made for the examples labeled “5”. It shows that the smoothed conformal predictor is not valid at the 95% confidence level on those examples, giving 11.7% of errors. Since the error rate of 5% is achieved on average, the error rate for some digits is better than 5%; for example, it is below 1% for the examples labeled “0”.

Inductive conformal transducers

Inductive conformal transducers, which output the p-values

$$\frac{|\{j : \alpha_j \geq \alpha_n\}|}{n - m_k}$$

(deterministic case) or

$$\frac{|\{j : \alpha_j > \alpha_n\}| + \tau_n |\{j : \alpha_j = \alpha_n\}|}{n - m_k}$$

(smoothed case), where

$$\alpha_j := A_{m_k+1}(\ell(x_1, y_1), \dots, (x_{m_k}, y_{m_k}), (x_j, y_j)), \quad j = m_k + 1, \dots, n$$

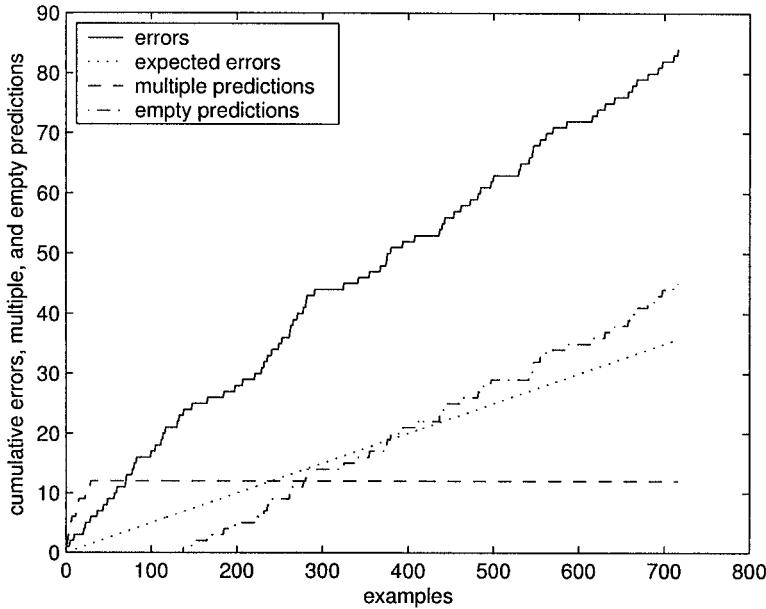


Fig. 4.6. The performance of the smoothed conformal predictor on the USPS data set for the examples labeled “5” at the 95% confidence level

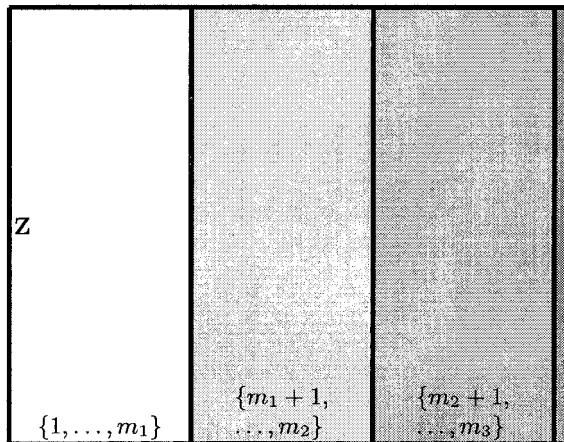


Fig. 4.7. Mondrian taxonomy corresponding to inductive conformal transducers

(cf. (4.2) and (4.3), p. 99), are also a special case of MCTs. The corresponding taxonomy is shown in Fig. 4.7. The result of §4.1 that ICPs are valid is a special case of Proposition 4.10. Similarly to conformal predictors, ICPs sometimes violate the property of label-wise validity.

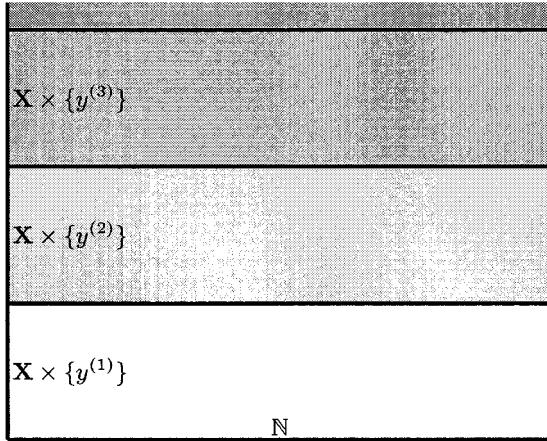


Fig. 4.8. Label-conditional Mondrian taxonomy

Label-conditional Mondrian conformal transducers

An important special case is where the category of an example is determined by its label. The corresponding taxonomy is shown in Fig. 4.8, where it is assumed that $\mathbf{Y} = \{y^{(1)}, \dots, y^{(L)}\}$.

Our experiments will be restricted to the “symmetric” case, where the same significance level (5%) is used for all categories. Figure 4.9 demonstrates empirically the category-wise validity of MCPs. In contrast to Fig. 4.6, the label-conditional MCP gives 5.3% of errors when the significance level is set to 5% for the label “5”. Figures 4.6 and 4.9 show that the correction in the number of errors results in an increased frequency of multiple predictions; there is also a decrease in the number of empty predictions.

Attribute-conditional Mondrian conformal transducers

The conditionality principle (Cox 1958b; Cox and Hinkley 1974, §2.3) is often illustrated using the following simple example (slightly modified) due to Cox (1958b). Suppose we have two instruments for measuring an unknown bit; at each trial one instrument is used once, and the instrument to use is chosen at random (tossing a fair coin). Instrument 1 is more accurate, with the probability of mistake equal to 1%, whereas the probability of mistake for instrument 2 is 5%. Formally, each object is a pair $x = (i, b)$, where $i \in \{1, 2\}$ is the instrument used and $b \in \{0, 1\}$ is the result of the measurement; the label $y \in \{0, 1\}$ is the true bit.

It is intuitively clear that at confidence level 99.5% the optimal valid confidence predictor (cf. the description of the Bayes confidence predictor in §3.4) will predict objects $(1, \dots)$ with singular predictions and will not predict objects $(2, \dots)$ at all (in the sense that its predictions will be the set $\{0, 1\}$ of

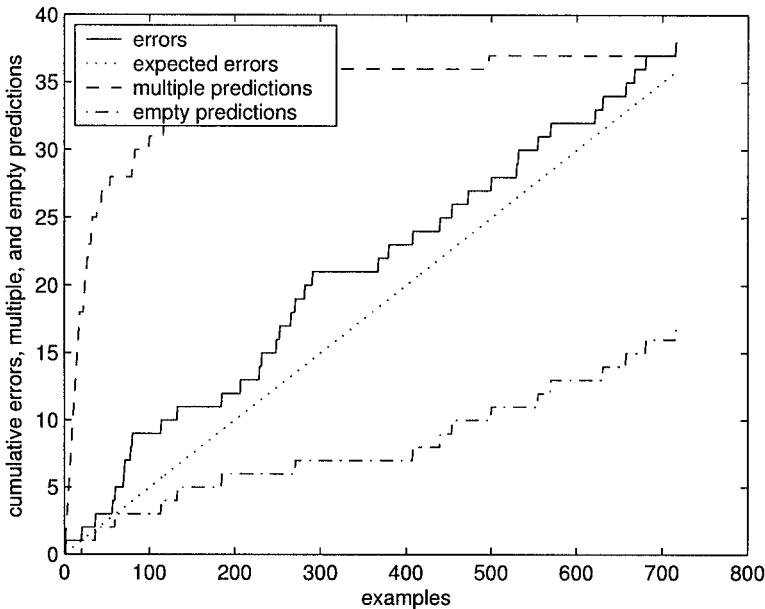


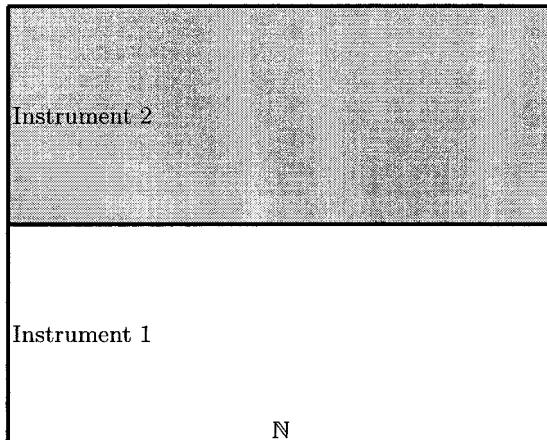
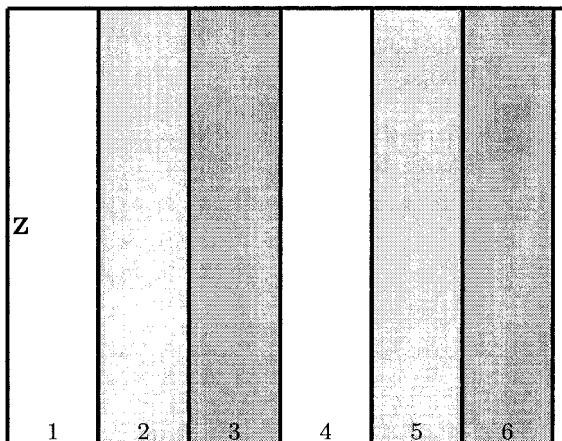
Fig. 4.9. The performance of the label-conditional MCP (based on the taxonomy $\kappa(n, (x, y)) = y$) on the USPS data set for the examples labeled as “5” at the 95% confidence level

all labels). At confidence level 97% the optimal valid confidence predictor will asymptotically predict all objects with singular predictions.

In both cases conditional validity is problematic (as argued by Cox); it does not prevent, however, the predictions from being valid on average. But the situation becomes even worse if we want to have two different significance levels for objects $(1, \dots)$ and $(2, \dots)$: if we take 0.5% for $(1, \dots)$ and 3% for $(2, \dots)$, any validity is lost.

The taxonomy for Cox’s example is shown in Fig. 4.10, where “Instrument 1” stands for the set of examples $((1, \dots), \dots)$ and “Instrument 2” stands for the set of examples $((2, \dots), \dots)$.

In our experiments with different data sets we have not seen as gross failures in the conformal predictor’s attribute-wise validity as those in the label-wise validity. The USPS data set does not have any natural attributes to condition on, since all attributes in it are of the same nature (the brightness level of a pixel) and continuous, but even for the data sets that do have natural attributes to condition on the conformal predictor’s conditional performance was reasonable.

**Fig. 4.10.** Cox's example**Fig. 4.11.** Slow teacher

Slow teacher

It is interesting that MCPs can be used to deal with the problem of slow teacher considered in §4.3 above. The delay l is assumed to be constant. Define $\kappa(n, z) := n \bmod (l + 1)$ (this is illustrated in Fig. 4.11 for $l = 2$) and take a nonconformity measure A_n (see (4.30) on p. 115) that depends on its arguments only via $\{z_j : j \in \{1, \dots, i-1, i+1, \dots, n\} \& \kappa_j = \kappa_n\}$ and (κ_n, z_i) . The corresponding smoothed MCP only needs a slow teacher with lag l , and Proposition 4.10 implies that it is not only asymptotically valid, but is valid in the sense that its errors are independent Bernoulli random variables with the right parameter. Of course, this predictor can only be used where there is a surfeit of examples.

4.6 Proofs

Proof of Theorem 4.2, I: $n_k/n_{k-1} \rightarrow 1$ is sufficient

We start from a simple general lemma about martingale differences.

Lemma 4.12. *If ξ_1, ξ_2, \dots is a martingale difference w.r. to σ -algebras $\mathcal{F}_1, \mathcal{F}_2, \dots$ and w_1, w_2, \dots is a sequence of positive numbers such that, for all $i = 1, 2, \dots$,*

$$\mathbb{E}(\xi_i^2 | \mathcal{F}_{i-1}) \leq w_i^2,$$

then

$$\mathbb{E}\left(\left(\frac{\xi_1 + \dots + \xi_n}{w_1 + \dots + w_n}\right)^2\right) \leq \frac{w_1^2 + \dots + w_n^2}{(w_1 + \dots + w_n)^2}.$$

Proof. Since elements of a martingale difference sequence are uncorrelated, we have

$$\mathbb{E}((\xi_1 + \dots + \xi_n)^2) = \sum_{1 \leq i \leq n} \mathbb{E}(\xi_i^2) + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}(\xi_i \xi_j) \leq \sum_{1 \leq i \leq n} w_i^2. \quad \square$$

Fix a significance level ϵ and a power probability distribution Q^∞ on \mathbf{Z}^∞ generating the examples $z_i = (x_i, y_i)$; the \mathcal{L} -taught smooth conformal predictor $\Gamma^\mathcal{L}$ is fed with the examples z_i and random numbers $\tau_i \in [0, 1]$. The error sequence and predictable error sequence of $\Gamma^\mathcal{L}$ will be denoted

$$e_n := \text{err}_n^\epsilon(\Gamma^\mathcal{L}) = \begin{cases} 1 & \text{if } y_n \notin \Gamma^{\mathcal{L}, \epsilon}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x_n, \tau_n) \\ 0 & \text{otherwise} \end{cases}$$

and

$$d_n := \overline{\text{err}}_n^\epsilon(\Gamma^\mathcal{L}) = (Q \times \mathbf{U}) \left\{ (x, y, \tau) \in \mathbf{Z} \times [0, 1] : \right. \\ \left. y \notin \Gamma^{\mathcal{L}, \epsilon}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau) \right\}.$$

Along with the original predictor $\Gamma^\mathcal{L}$ we also consider the *ghost predictor*, which is Γ fed with the examples

$$z'_1 = (x'_1, y'_1) := z_{\mathcal{L}(n_1)}, z'_2 = (x'_2, y'_2) := z_{\mathcal{L}(n_2)}, \dots$$

and random numbers τ'_1, τ'_2, \dots (independent from each other and from the sequences z_i and τ_i). The ghost predictor is given all labels and each label is given without delay. Notice that its input sequence $z_{\mathcal{L}(n_1)}, z_{\mathcal{L}(n_2)}, \dots$ is also distributed according to Q^∞ . The error and predictable error sequences of the ghost predictor are

$$\begin{aligned} e'_n &:= \text{err}_n^\epsilon(\Gamma, (z'_1, z'_2, \dots)) \\ &= \begin{cases} 1 & \text{if } y'_n \notin \Gamma^\epsilon(x'_1, \tau'_1, y'_1, \dots, x'_{n-1}, \tau'_{n-1}, y'_{n-1}, x'_n, \tau'_n) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and

$$\begin{aligned} d'_n &:= \overline{\text{err}}_n^\epsilon(\Gamma, (z'_1, z'_2, \dots)) = (Q \times \mathbf{U}) \left\{ (x, y, \tau) \in \mathbf{Z} \times [0, 1] : \right. \\ &\quad \left. y \notin \Gamma^\epsilon(x'_1, \tau'_1, y'_1, \dots, x'_{n-1}, \tau'_{n-1}, y'_{n-1}, x, \tau) \right\}. \end{aligned}$$

It is clear that, for each k , d_n is the same for all $n = n_{k-1} + 1, \dots, n_k$, their common value being

$$d_{n_k} = d'_k. \quad (4.31)$$

Corollary 4.13. *For each k ,*

$$\begin{aligned} \mathbb{E} \left(\left(\frac{(e'_1 - \epsilon)n_1 + (e'_2 - \epsilon)(n_2 - n_1) + \dots + (e'_k - \epsilon)(n_k - n_{k-1})}{n_k} \right)^2 \right) \\ \leq \frac{n_1^2 + (n_2 - n_1)^2 + \dots + (n_k - n_{k-1})^2}{n_k^2}. \end{aligned}$$

Proof. It is sufficient to apply Lemma 4.12 to $w_i := n_i - n_{i-1}$ (n_0 is understood to be 0 in this section), the independent zero-mean (by Proposition 2.4 on p. 27) random variables $\xi_i := (e'_i - \epsilon)w_i$, and the σ -algebras \mathcal{F}_i generated by ξ_1, \dots, ξ_i . \square

Corollary 4.14. *For each k ,*

$$\begin{aligned} \mathbb{E} \left(\left(\frac{(e'_1 - d'_1)n_1 + (e'_2 - d'_2)(n_2 - n_1) + \dots + (e'_k - d'_k)(n_k - n_{k-1})}{n_k} \right)^2 \right) \\ \leq \frac{n_1^2 + (n_2 - n_1)^2 + \dots + (n_k - n_{k-1})^2}{n_k^2}. \end{aligned}$$

Proof. Use Lemma 4.12 for $w_i := n_i - n_{i-1}$, $\xi_i := (e'_i - d'_i)w_i$, and the σ -algebras \mathcal{F}_i generated by z'_1, \dots, z'_i and τ'_1, \dots, τ'_i . \square

Corollary 4.15. *For each k ,*

$$\mathbb{E} \left(\frac{(e_1 - d_1) + (e_2 - d_2) + \dots + (e_{n_k} - d_{n_k})}{n_k} \right)^2 \leq \frac{1}{n_k}.$$

Proof. Apply Lemma 4.12 to $w_i := 1$, $\xi_i := e_i - d_i$, and the σ -algebras \mathcal{F}_i generated by z_1, \dots, z_i and τ_1, \dots, τ_i . \square

Lemma 4.16. *If $\lim_{k \rightarrow \infty} (n_k/n_{k-1}) = 1$ for some strictly increasing sequence of positive integers n_1, n_2, \dots , then*

$$\lim_{k \rightarrow \infty} \frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} = 0.$$

Proof. For any $\delta > 0$, there exists a K such that $\frac{n_k - n_{k-1}}{n_{k-1}} < \delta$ for any $k > K$. Therefore,

$$\begin{aligned} & \frac{n_1^2 + (n_2 - n_1)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} \\ & \leq \frac{n_K^2}{n_k^2} + \frac{(n_{K+1} - n_K)^2 + \cdots + (n_k - n_{k-1})^2}{n_k^2} \\ & \leq \frac{n_K^2}{n_k^2} + \frac{n_{K+1} - n_K}{n_K} \frac{n_{K+1} - n_K}{n_k} + \frac{n_{K+2} - n_{K+1}}{n_{K+1}} \frac{n_{K+2} - n_{K+1}}{n_k} + \cdots \\ & + \frac{n_k - n_{k-1}}{n_{k-1}} \frac{n_k - n_{k-1}}{n_k} \leq \frac{n_K^2}{n_k^2} + \delta \frac{(n_{K+1} - n_K) + \cdots + (n_k - n_{k-1})}{n_k} \leq 2\delta \end{aligned}$$

from some k on. \square

Now it is easy to finish the proof of the first part of the theorem. In combination with Chebyshev's inequality and Lemma 4.16, Corollary 4.13 implies that

$$\frac{(e'_1 - \epsilon)n_1 + (e'_2 - \epsilon)(n_2 - n_1) + \cdots + (e'_k - \epsilon)(n_k - n_{k-1})}{n_k} \rightarrow 0$$

in probability; using the notation $k(i) := \min\{k : n_k \geq i\} = s(i) + 1$, we can rewrite this as

$$\frac{1}{n_k} \sum_{i=1}^{n_k} (e'_{k(i)} - \epsilon) \rightarrow 0. \quad (4.32)$$

Similarly, (4.31) and Corollary 4.14 imply

$$\frac{1}{n_k} \sum_{i=1}^{n_k} (e'_{k(i)} - d'_{k(i)}) = \frac{1}{n_k} \sum_{i=1}^{n_k} (e'_{k(i)} - d_i) \rightarrow 0, \quad (4.33)$$

and Corollary 4.15 implies

$$\frac{1}{n_k} \sum_{i=1}^{n_k} (e_i - d_i) \rightarrow 0 \quad (4.34)$$

(all convergences are in probability). Combining (4.32)–(4.34), we obtain

$$\frac{1}{n_k} \sum_{i=1}^{n_k} (e_i - \epsilon) \rightarrow 0; \quad (4.35)$$

the condition $n_k/n_{k-1} \rightarrow 1$ allows us to replace n_k with n in (4.35).

Proof of Theorem 4.2, II: $n_k/n_{k-1} \rightarrow 1$ is necessary

Fix $\epsilon := 5\%$. As a first step, we construct the example space \mathbf{Z} , the probability distribution Q on \mathbf{Z} and a smoothed conformal predictor for which d'_k deviate consistently from ϵ . Let $\mathbf{X} = \{0\}$, $\mathbf{Y} = \{0, 1\}$, so that z_i is, essentially, always 0 or 1. The probability distribution Q is uniform on \mathbf{Z} : $Q\{0\} = Q\{1\} = 1/2$. The nonconformity measure is

$$\alpha_i = A(\{\zeta_1, \dots, \zeta_{i-1}, \zeta_{i+1}, \dots, \zeta_k\}, \zeta_i) := \begin{cases} \zeta_i & \text{if } \zeta_1 + \dots + \zeta_k \text{ is even} \\ 1 - \zeta_i & \text{if } \zeta_1 + \dots + \zeta_k \text{ is odd} \end{cases}.$$

It follows from the central limit theorem that

$$\frac{|\{i = 1, \dots, k : z'_i = 1\}|}{k} \in (0.4, 0.6) \quad (4.36)$$

with probability at least 99% for k large enough. We will show that d'_k deviates significantly from ϵ with probability at least 99% for sufficiently large k . Let $\alpha_i := A(\{z'_1, \dots, z'_{i-1}, z'_{i+1}, \dots, z'_k\}, z'_i)$ with z'_k is interpreted as y (corresponding to (x, y) in the previous subsection). There are two possibilities:

- If $z'_1 + \dots + z'_{k-1}$ is odd, then

$$\begin{aligned} z'_k = 1 &\implies z'_1 + \dots + z'_{k-1} + z'_k \text{ is even} \implies \alpha_k = z'_k = 1 \\ z'_k = 0 &\implies z'_1 + \dots + z'_{k-1} + z'_k \text{ is odd} \implies \alpha_k = 1 - z'_k = 1. \end{aligned}$$

In both cases we have $\alpha_k = 1$ and, therefore, outside an event of probability at most 1%,

$$\begin{aligned} d'_k &= (Q \times \mathbf{U}) \{(y, \tau) : \tau |\{i = 1, \dots, k : \alpha_i = 1\}| \leq k\epsilon\} \\ &= \int_{\mathbf{Y}} 1 \wedge \frac{k\epsilon}{|\{i = 1, \dots, k : \alpha_i = 1\}|} Q(dy) \geq \frac{k\epsilon}{0.7k} = \frac{10}{7}\epsilon. \end{aligned}$$

- If $z'_1 + \dots + z'_{k-1}$ is even, then

$$\begin{aligned} z'_k = 1 &\implies z'_1 + \dots + z'_{k-1} + z'_k \text{ is odd} \implies \alpha_k = 1 - z'_k = 0 \\ z'_k = 0 &\implies z'_1 + \dots + z'_{k-1} + z'_k \text{ is even} \implies \alpha_k = z'_k = 0. \end{aligned}$$

In both cases $\alpha_k = 0$ and, therefore, outside an even of probability at most 1%,

$$\begin{aligned} d'_k &= (Q \times \mathbf{U}) \left\{ (y, \tau) : \right. \\ &\quad \left. |\{i = 1, \dots, k : \alpha_i = 1\}| + \tau |\{i = 1, \dots, k : \alpha_i = 0\}| \leq k\epsilon \right\} \\ &\leq (Q \times \mathbf{U}) \{(y, \tau) : 0.3k \leq k\epsilon\} = 0. \end{aligned}$$

To summarize, for large enough k ,

$$|d'_k - \epsilon| = |d_{n_k} - \epsilon| > \epsilon/3 \quad (4.37)$$

with probability at least 99% (cf. (4.31); we write 99% rather than 98% since the two exceptional events of probability 1% coincide: both are the complement of (4.36)).

Suppose that

$$\frac{1}{n} \sum_{i=1}^n e_i \rightarrow \epsilon \quad (4.38)$$

in probability; we will deduce that $n_k/n_{k-1} \rightarrow 1$. By (4.34) (remember that Corollary 4.15 and, therefore, (4.34) do not depend on the condition $n_k/n_{k-1} \rightarrow 1$) and (4.38) we have

$$\frac{1}{n_k} \sum_{i=1}^{n_k} d_i \rightarrow \epsilon ;$$

we can rewrite this in the form

$$\sum_{i=1}^{n_k} d_i = n_k(\epsilon + o(1))$$

(all $o(1)$ are in probability). This equality implies

$$\sum_{k=0}^K d_{n_k} (n_k - n_{k-1}) = n_K(\epsilon + o(1))$$

and

$$\sum_{k=0}^{K-1} d_{n_k} (n_k - n_{k-1}) = n_{K-1}(\epsilon + o(1)) ;$$

subtracting the last equality from the penultimate one we obtain

$$d_{n_K} (n_K - n_{K-1}) = (n_K - n_{K-1})\epsilon + o(n_K) ,$$

i.e.,

$$(d_{n_K} - \epsilon) (n_K - n_{K-1}) = o(n_K) .$$

In combination with (4.37), this implies $n_K - n_{K-1} = o(n_K)$, i.e., $n_K/n_{K-1} \rightarrow 1$ as $K \rightarrow \infty$.

Proof of Theorem 4.4

This proof is similar to the proof of Theorem 4.2. (The definition of $\Gamma^{\mathcal{L}}$ being asymptotically exact involves the assumption of exchangeability rather than randomness; however, since we assumed that \mathbf{Z} is a Borel space, de Finetti's theorem, stated in §A.5, shows that these assumptions are equivalent in our current context.) Instead of Corollaries 4.13, 4.14, and 4.15 we now have:

Corollary 4.17. As $k \rightarrow \infty$,

$$\frac{(e'_1 - \epsilon)n_1 + (e'_2 - \epsilon)(n_2 - n_1) + \cdots + (e'_k - \epsilon)(n_k - n_{k-1})}{n_k} \rightarrow 0 \quad a.s.$$

Proof. It is sufficient to apply Kolmogorov's strong law of large numbers (stated in §A.6) to the independent zero-mean random variables $\xi_i = (e'_i - \epsilon)(n_i - n_{i-1})$. Condition (A.8) (p. 286) follows from

$$\sum_{i=1}^{\infty} \frac{(n_i - n_{i-1})^2}{n_i^2} < \infty,$$

which is equivalent to (4.18). \square

Corollary 4.18. As $k \rightarrow \infty$,

$$\frac{(e'_1 - d'_1)n_1 + (e'_2 - d'_2)(n_2 - n_1) + \cdots + (e'_k - d'_k)(n_k - n_{k-1})}{n_k} \rightarrow 0 \quad a.s.$$

Proof. Apply the martingale strong law of large numbers (§A.6) to the martingale difference $\xi_i = (e'_i - d'_i)(n_i - n_{i-1})$ w.r. to the σ -algebras \mathcal{F}_i generated by z'_1, \dots, z'_i and τ'_1, \dots, τ'_i . \square

Corollary 4.19. As $k \rightarrow \infty$,

$$\frac{(e_1 - d_1) + (e_2 - d_2) + \cdots + (e_{n_k} - d_{n_k})}{n_k} \rightarrow 0 \quad a.s.$$

Proof. Apply the martingale strong law of large numbers to the martingale difference $\xi_i = e_i - d_i$ w.r. to the σ -algebras \mathcal{F}_i generated by z_1, \dots, z_i and τ_1, \dots, τ_i . \square

Corollary 4.17 can be rewritten as (4.32), Corollary 4.18 as (4.33), and Corollary 4.19 as (4.34); all convergences are now almost certain. Combining (4.32)–(4.34), we obtain (4.35). It remains to replace n_k with n , as before.

Proof of Theorem 4.8

We will only consider the case where Γ is a smoothed conformal predictor (the proof for deterministic Γ is almost identical: just ignore all random numbers τ).

Fix a significance level ϵ and define

$$\overline{\text{mult}}_n(\Gamma^{\mathcal{L}}) = (Q \times \mathbf{U}) \left\{ (x, \tau) \in \mathbf{X} \times [0, 1] : \right. \\ \left. |\Gamma^{\mathcal{L}, \epsilon}(x_1, \tau_1, y_1, \dots, x_{n-1}, \tau_{n-1}, y_{n-1}, x, \tau)| > 1 \right\},$$

$$\overline{\text{mult}}_k(\Gamma) = (Q \times \mathbf{U}) \left\{ (x, \tau) \in \mathbf{X} \times [0, 1] : \right. \\ \left. |\Gamma^\epsilon(x'_1, \tau'_1, y'_1, \dots, x'_{k-1}, \tau'_{k-1}, y'_{k-1}, x, \tau)| > 1 \right\},$$

$$\overline{\text{Mult}}_n(\Gamma^{\mathcal{L}}) = \sum_{i=1}^n \overline{\text{mult}}_i(\Gamma^{\mathcal{L}}), \quad \overline{\text{Mult}}_k(\Gamma) = \sum_{i=1}^k \overline{\text{mult}}_i(\Gamma).$$

Since $\text{Mult}_n^\epsilon(\Gamma^{\mathcal{L}}) - \overline{\text{Mult}}_n(\Gamma^{\mathcal{L}})$ is a martingale and

$$|\text{mult}_n^\epsilon(\Gamma^{\mathcal{L}}) - \overline{\text{mult}}_n(\Gamma^{\mathcal{L}})| \leq 1,$$

the martingale strong law of large numbers (see §A.6) implies that

$$\lim_{n \rightarrow \infty} \frac{\text{Mult}_n^\epsilon(\Gamma^{\mathcal{L}}) - \overline{\text{Mult}}_n(\Gamma^{\mathcal{L}})}{n} = 0 \quad \text{a.s.} \quad (4.39)$$

Analogously,

$$\lim_{k \rightarrow \infty} \frac{\text{Mult}_k^\epsilon(\Gamma, (z'_1, z'_2, \dots)) - \overline{\text{Mult}}_k(\Gamma)}{k} = 0 \quad \text{a.s.} \quad (4.40)$$

By (4.39) and (4.40), we can replace Mult^ϵ with $\overline{\text{Mult}}$ in the definitions of $U^\epsilon(\Gamma^{\mathcal{L}}, Q)$ and $U^\epsilon(\Gamma, Q)$.

It is clear that

$$\overline{\text{mult}}_n(\Gamma^{\mathcal{L}}) = \overline{\text{mult}}_{k(n)}(\Gamma)$$

for all n . Combining this with $k(n) = n/c + O(1)$, we obtain

$$\sum_{i=1}^n \overline{\text{mult}}_i(\Gamma^{\mathcal{L}}) = c \sum_{i=1}^{\lfloor n/c \rfloor} \overline{\text{mult}}_i(\Gamma) + O(1),$$

and so $\overline{\text{Mult}}_n(\Gamma^{\mathcal{L}}) = c \overline{\text{Mult}}_{\lfloor n/c \rfloor}(\Gamma) + o(n)$. The statement of the theorem immediately follows.

4.7 Bibliographical remarks

Computationally efficient hedged prediction

To cope with the relative computational inefficiency of conformal predictors, inductive conformal predictors were introduced in Papadopoulos et al. 2002a and Papadopoulos et al. 2002b in the off-line setting and in Vovk 2002b in the on-line setting. Before the appearance of inductive conformal predictors, several other possibilities had been studied, such as “competitive transduction” (Saunders 2000) and “transduction with hashing” (Saunders et al. 2000; Saunders 2000).

Specific learning algorithms and nonconformity measures

Equation (4.11) is sometimes known as the Sherman–Morrison formula (it can be checked easily by multiplying the right-hand side by $K + uv'$ on the left and simplifying); for details, see Henderson and Searle 1981.

The bootstrap was proposed by Efron (1979). For recent reviews, see Efron 2003 and other articles in the same issue of *Statistical Science*. There are two main varieties of regression bootstrap: “bootstrapping residuals” and “bootstrapping cases”. (For details, see Montgomery et al. 2001, pp. 509–510, or Draper and Smith 1998, pp. 285–286.) We only gave an example of using the first of these procedures, following Davison and Hinkley 1997 (Algorithm 6.4), the original idea being due to Stine (1985).

Decision trees are reviewed, besides Mitchell 1997 (Chap. 3), in Ripley 1996 (Chap. 7); the latter contains many pointers to the relevant literature. The C4.5 algorithm was introduced in Quinlan 1993.

In our description of hedged prediction based on boosting we followed Proedrou 2003; both definitions (4.14) and (4.15) of conformity scores are due to him. The first boosting algorithm was proposed by Schapire (1990); AdaBoost is due to Freund and Schapire (1997).

Neural networks are popular in both classification and regression; good references are Bishop 1995 and Ripley 1996. The current wave of interest was mainly initiated by Rumelhart and McClelland (1986).

Cox 1970 and more recent Hosmer and Lemeshow 2000 are useful sources for logistic regression. Cox 1958a may be the first publication describing logistic regression, although Jerome Cornfield might have used it several years before 1958 (Reid 1994, p. 448).

In this book we have given examples of nonconformity measures based on least squares, ridge regression, logistic regression, nearest neighbors, support vector machines, decision trees, boosting, bootstrap, and neural networks. The number of known machine learning algorithms is huge, however, and potentially any of them can be used as a source of nonconformity measures; in particular, we did not touch the important class of genetic algorithms (see, e.g., Mitchell 1996). For a very readable introduction to machine learning algorithms, see Mitchell 1997, and for recent developments see the proceedings of the numerous machine learning conferences, such as NIPS, ICML, UAI, COLT, and ALT.

Weak teachers

The characterization of lazy teachers for which conformal prediction is valid in probability was obtained by Ilia Nouretdinov. The general notion of a teaching schedule and the device of a “ghost predictor” is due to Daniil Ryabko (Ryabko et al. 2003). Nouretdinov’s result (our Theorem 4.2), generalized in light of Ryabko et al. 2003, appeared in Nouretdinov and Vovk 2003. Theorems 4.4 and 4.8 are from Ryabko et al. 2003.

Mondrian conformal predictors

For further information, see Vovk et al. 2003a.

5

Probabilistic prediction I: impossibility results

In this and following chapters we will discuss probabilistic prediction under unconstrained randomness. As we noticed in Chap 1, there are several possible goals associated with probabilistic prediction; some of these are attainable and some are not. This chapter concentrates on unattainable goals (the following chapter will also have some negative results, but they will play an auxiliary role: to help us state interesting attainable goals).

Two important factors that determine the feasibility of probabilistic prediction are:

- Are we interested in asymptotic results or are we interested in the finite world of our experience?
- Do we want to estimate the true probabilities or do we just want some numbers that can pass for probabilities?

In some parts of this book (namely, when discussing efficiency of conformal predictors in Chap. 3 and of Venn predictors in the following chapter) we are interested in asymptotic results, but in this chapter we will emphasize the limited interest of such results for practical learning problems. The main part of this chapter is about estimating the true probabilities, but in the last section we also state and prove a result showing that already in the simplest problem of probabilistic prediction it is not possible to produce numbers that can pass for probabilities, if “can pass” is understood in the sense of the algorithmic theory of randomness.

In §§5.1–5.2 we state the main negative result of this chapter: probabilistic prediction in the sense of estimating true probabilities from a given finite training set is impossible under unconstrained randomness unless one can use precise repetition of objects in the training set. As we explained in Chap. 1, we are primarily interested in learning methods that work in high-dimensional environment, where precise repetitions are hardly possible. In §5.1 we briefly discuss the nature of the assumption of no repetitions, and in §5.2 state the mathematical result.

To state our results in their strongest form, in this chapter we use the assumption of randomness (rather than exchangeability): the examples z_1, z_2, \dots are generated from a power distribution Q^∞ on \mathbf{Z}^∞ .

5.1 Diverse data sets

This section continues the discussion of learning under unconstrained randomness started in Chap. 1.

Important features of even moderately interesting machine-learning problems (such as hand-written digit recognition) are:

1. It might be reasonable to assume that different objects are drawn from the same probability distribution independently of each other, but we cannot make any further assumptions beyond randomness.
2. The objects we are presented with will typically have a fairly complicated structure (in the case of the USPS data set, every object is a 16×16 gray-scale matrix).
3. In general, we do not expect the objects to be repeated *precisely*.

Features 2 and 3 are closely connected but still different: one can imagine even complicated patterns repeated precisely (such as two twins' genetic code), and in many cases one can expect that even simple unstructured objects, such as real numbers, are never repeated (if they, e.g., are generated by a continuous probability distribution).

Of course, no learning method can work unless the probability distribution generating the data is benign in some respects; for example, different instances of the same digit in the USPS data set look reasonably similar. But we may say that a type of learning problems is infeasible under unconstrained randomness if those problems can be solved *only* in the case where the data set has repeated objects.

In some cases estimation of probabilities is possible: for example, if objects are absent ($|\mathbf{X}| = 1$) and $\mathbf{Y} = \{0, 1\}$, estimation of the probability that $y_n = 1$ given y_1, \dots, y_{n-1} is easy (this was one of the first problems solved by the mathematical theory of probability; see Chap. 10). But we will see that in general the problem of estimation of probabilities should be classified as infeasible in learning under unconstrained randomness.

5.2 Impossibility of estimation of probabilities

The prediction problem considered in this chapter is more challenging than that of the preceding chapters in that the prediction algorithm is required not just to predict the next label y_n but to estimate the conditional probabilities $Q_{\mathbf{Y}|\mathbf{X}}(y | x_n)$ for $y \in \mathbf{Y}$. The label space \mathbf{Y} will be assumed to be finite, and we start from the simplest binary case, $\mathbf{Y} = \{0, 1\}$. We will also assume that

the object space \mathbf{X} is finite. This is not a restrictive assumption from the practical point of view (all data sets we are aware of allocate a fixed number of bits for each objects) but will allow us to avoid complications coming from the foundations of probability.

Binary case

Suppose $\mathbf{Y} = \{0, 1\}$; in this case the estimation of probabilities $Q_{\mathbf{Y}|\mathbf{X}}(y | x_n)$, $y \in \mathbf{Y}$, boils down to estimating $Q_{\mathbf{Y}|\mathbf{X}}(1 | x_n)$. The notions introduced in this subsection will be redefined (essentially, generalized) in the next one.

A *probability estimator* is a measurable family of functions

$$\Gamma^\epsilon : (x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \mapsto A, \quad (5.1)$$

where the significance level ϵ ranges over $(0, 1)$, n ranges over the positive integers, and A over subsets of the interval $[0, 1]$ (typically A will be an interval, open, closed, or mixed), which satisfies, for all n , all incomplete data sequences, and all significance levels $\epsilon_1 > \epsilon_2$,

$$\Gamma^{\epsilon_1}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma^{\epsilon_2}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n). \quad (5.2)$$

The measurability of (5.1) means that the set

$$\{(\epsilon, x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, p) : p \in \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)\} \quad (5.3)$$

is a measurable subset of $(0, 1) \times \mathbf{Z}^{n-1} \times \mathbf{X} \times [0, 1]$ for all $n = 1, 2, \dots$.

We say that such a probability estimator is *weakly valid* if, for any probability distribution Q on \mathbf{Z} ,

$$\begin{aligned} Q^\infty \{(x_1, y_1, x_2, y_2, \dots) : Q_{\mathbf{Y}|\mathbf{X}}(1 | x_n) \in \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)\} \\ \geq 1 - \epsilon. \end{aligned}$$

In other words, a probability estimator should cover the true conditional probability of 1 with probability at least $1 - \epsilon$, under any power probability distribution generating the data. The word “weakly” is to emphasize the lack of the requirement of independence of errors at different trials (we say that a probability estimator makes an error if it does not cover the true probability). The following result is our formalization of the impossibility of estimation of probabilities in the binary case.

Proposition 5.1. *For any weakly valid probability estimator Γ there is another weakly valid probability estimator $\tilde{\Gamma}$ such that, for any incomplete data sequence x_1, y_1, \dots, x_n that does not contain repeated objects and for any $\epsilon \in (0, 1)$,*

$$\begin{aligned} \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) &\subseteq (0, 1) \\ \implies \tilde{\Gamma}^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) &= \emptyset. \end{aligned} \quad (5.4)$$

Let us see why the conclusion (5.4) can be interpreted as saying that nontrivial probability estimation is impossible. Let us assume for simplicity that the region output by Γ is an interval $[a, b]$. The case where $[a, b]$ contains 0 or 1 corresponds to classification: for example, if $a = 0$, the prediction $[a, b] = [0, b]$ essentially says that we expect the new label to be 0, with b quantifying our confidence in this prediction. Genuine probability estimation corresponds to the case $[a, b] \subseteq (0, 1)$, and the theorem says that if we can output such an estimate, we can also output an empty (which is better, because more precise) estimate. An empty estimate is analogous to a contradiction in logic: the foundation for the inference is not sound (the examples are not typical under the randomness assumption) and everything can be deduced. Slightly abusing the standard statistical terminology, we may say that an estimate $[a, b] \subseteq (0, 1)$ is not admissible: it can be improved to being false.

Multi-label case

Let now \mathbf{Y} be an arbitrary finite set (with the discrete σ -algebra). The notions defined in the binary case in the previous subsection can be carried over to the general case as follows. A *probability estimator* is a family of functions (5.1), where $\epsilon \in (0, 1)$, $n = 0, 1, \dots$, and A ranges over subsets of the family $\mathbf{P}(\mathbf{Y})$ of all probability distributions on \mathbf{Y} , which is required to satisfy the conditions of consistency (5.2) (whenever $\epsilon_1 \geq \epsilon_2$) and measurability (i.e., (5.3) being a measurable subset of $(0, 1) \times \mathbf{Z}^{n-1} \times \mathbf{X} \times \mathbf{P}(\mathbf{Y})$ for all n). Such a probability estimator is *weakly valid* if, for any probability distribution Q on \mathbf{Z} , any n , and any incomplete data sequence x_1, y_1, \dots, x_n ,

$$Q^\infty \{(x_1, y_1, x_2, y_2, \dots) : Q(\cdot | x_n) \in \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)\} \geq 1 - \epsilon,$$

where $Q(\cdot | x_n)$ is the probability distribution on \mathbf{Y} assigning probability $Q_{\mathbf{Y}|\mathbf{X}}(y | x_n)$ to each $y \in \mathbf{Y}$.

Let $\mathbf{P}^\circ(\mathbf{Y})$ be the subset of $\mathbf{P}(\mathbf{Y})$ consisting of the *non-degenerate* probability distributions p on \mathbf{Y} , i.e., those distributions p that satisfy $\max_{y \in \mathbf{Y}} p(y) < 1$. The following theorem is a generalization of Proposition 5.1.

Theorem 5.2. *For any weakly valid probability estimator Γ there is another weakly valid probability estimator $\tilde{\Gamma}$ such that, for any incomplete data sequence x_1, y_1, \dots, x_n that does not contain repeated objects and for any $\epsilon \in (0, 1)$,*

$$\begin{aligned} \Gamma^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) &\subseteq \mathbf{P}^\circ(\mathbf{Y}) \\ \implies \tilde{\Gamma}^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) &= \emptyset. \end{aligned} \quad (5.5)$$

5.3 Proof of Theorem 5.2

We first sketch the idea behind the proof. Let Q be the true probability distribution on \mathbf{Z} generating the individual examples. We can imagine that the incomplete data sequence (x_1, y_1, \dots, x_n) , supposed not to contain repeated objects, is generated in two steps. First for each $x \in \mathbf{X}$ we choose randomly $g(x) \in \mathbf{Y}$ setting $g(x) := y$ with probability $Q_{\mathbf{Y}|\mathbf{X}}(y|x)$; this is done for each x independently of the other xs . After such a function $g : \mathbf{X} \rightarrow \mathbf{Y}$ is generated, we generate $x_i \in \mathbf{X}$, $i = 1, \dots, n$, independently from $Q_{\mathbf{X}}$ and finally set $y_i := g(x_i)$.

There is no way to tell from the data sequence x_1, y_1, \dots, x_n whether it was generated from Q or from $Q_{\mathbf{X}}$ and g and so, even if the true distribution $Q_{\mathbf{Y}|\mathbf{X}}(y|x_n)$, $y \in \mathbf{Y}$, is not degenerate, we will not be able to exclude the corners of the simplex $\mathbf{P}(\mathbf{Y})$ from our forecast $\Gamma^\epsilon(x_1, y_1, \dots, x_n)$ (unless the sequence x_1, y_1, \dots, x_n itself is untypical of Q).

Probability estimators and statistical tests

Our first step will be to reduce Theorem 5.2 to a statement about “incomplete statistical tests”. Complete statistical tests, which are not really needed in this book but clarify the notion of incomplete statistical tests, will be discussed in the next subsection. In this chapter statistical testing serves merely as a technical tool; it will be discussed more systematically in the following two chapters.

An *incomplete statistical test* is a measurable function $t : \mathbf{Z}^* \times \mathbf{X} \times \mathbf{P}(\mathbf{Z}) \rightarrow [0, 1]$ such that, for any n , $Q \in \mathbf{P}(\mathbf{Z})$, and $\epsilon \in [0, 1]$:

$$Q^\infty\{(x_1, y_1, x_2, y_2, \dots) : t_Q(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \leq \epsilon\} \leq \epsilon$$

(we write Q as a lower index: $t_Q(x_1, y_1, \dots, x_n)$ instead of $t(x_1, y_1, \dots, x_n, Q)$). If $t_Q(x_1, y_1, \dots, x_n) \leq \epsilon$, we say that the test t *rejects* Q at level ϵ (the incomplete data sequence x_1, y_1, \dots, x_n will be clear from the context).

There is a close connection (although by no means equivalence) between weakly valid probability estimators and incomplete statistical tests:

- If Γ is a weakly valid probability estimator,

$$t_Q(x_1, y_1, \dots, x_n) := \sup \{\epsilon : Q_{\mathbf{Y}|\mathbf{X}}(\cdot | x_n) \in \Gamma^\epsilon(x_1, y_1, \dots, x_n)\} \quad (5.6)$$

(with $\sup \emptyset := 0$) is an incomplete statistical test.

- If t is an incomplete statistical test,

$$\begin{aligned} \Gamma^\epsilon(x_1, y_1, \dots, x_n) &:= \{p \in \mathbf{P}(\mathbf{Y}) : \\ &\exists Q \in \mathbf{P}(\mathbf{Z}) : t_Q(x_1, y_1, \dots, x_n) > \epsilon \text{ \& } p = Q_{\mathbf{Y}|\mathbf{X}}(\cdot | x_n)\} \end{aligned} \quad (5.7)$$

is a weakly valid probability estimator.

The second statement is obvious, and the first follows from

$$\begin{aligned} Q^\infty\{(x_1, y_1, x_2, y_2, \dots) : t_Q(x_1, y_1, \dots, x_n) \leq \epsilon_0\} \\ = Q^\infty\{(x_1, y_1, x_2, y_2, \dots) : \forall \epsilon > \epsilon_0 : Q_{\mathbf{Y}|\mathbf{X}}(\cdot | x_n) \notin \Gamma^\epsilon(x_1, y_1, \dots, x_n)\} \\ \leq \inf_{\epsilon > \epsilon_0} \epsilon = \epsilon_0 . \end{aligned}$$

Complete statistical tests

To help the reader's intuition, we briefly discuss a more natural notion of statistical test. A function $t : \mathbf{Z}^* \times \mathbf{P}(\mathbf{Z}) \rightarrow [0, 1]$ is called a *complete statistical test* if, for any n , $Q \in \mathbf{P}(\mathbf{Z})$, and $\epsilon \in [0, 1]$:

$$Q^\infty\{(z_1, z_2, \dots) : t_Q(z_1, \dots, z_n) \leq \epsilon\} \leq \epsilon .$$

Complete statistical tests provide a means of testing whether a data sequence z_1, \dots, z_n could have been generated from a power distribution Q^n .

To any incomplete statistical test $t_Q(x_1, y_1, \dots, x_n)$ corresponds the complete statistical test

$$\tilde{t}_Q(x_1, y_1, \dots, x_n, y_n) := t_Q(x_1, y_1, \dots, x_n)$$

and to any complete statistical test $t_Q(x_1, y_1, \dots, x_n, y_n)$ corresponds the incomplete statistical test

$$\tilde{t}_Q(x_1, y_1, \dots, x_n) := \max_{y \in \mathbf{Y}} t_Q(x_1, y_1, \dots, x_n, y) .$$

Restatement of the theorem in terms of statistical tests

Now we can start implementing the idea of the proof sketched above. For any probability distribution $Q \in \mathbf{P}(\mathbf{Z})$ and any $g : \mathbf{X} \rightarrow \mathbf{Y}$ define the weight

$$\pi_Q(g) := \prod_{x \in \mathbf{X}} Q_{\mathbf{Y}|\mathbf{X}}(g(x) | x)$$

(this is the probability of generating g according to the procedure in the proof sketch) and define the probability distribution Q_g on \mathbf{Z} by the conditions that $(Q_g)_\mathbf{X} = Q_\mathbf{X}$ and that $(Q_g)_{\mathbf{Y}|\mathbf{X}}(\cdot | x)$ is concentrated at the point $g(x)$ for all $x \in \mathbf{X}$ (Q_g is the probability distribution governing the second stage of the procedure of generating the incomplete data sequence in the proof sketch). Notice that

$$\sum_{g: \mathbf{X} \rightarrow \mathbf{Y}} \pi_Q(g) = 1 .$$

Theorem 5.2 will be deduced from the following statement.

Proposition 5.3. *For any incomplete statistical test t there exists an incomplete statistical test T such that the following holds. If (x_1, y_1, \dots, x_n) is an incomplete data sequence without repeated objects and $Q \in \mathbf{P}(\mathbf{Z})$, then there exists $g : \mathbf{X} \rightarrow \mathbf{Y}$ such that*

$$T_Q(x_1, y_1, \dots, x_n) \leq t_{Q_g}(x_1, y_1, \dots, x_n). \quad (5.8)$$

To prove Proposition 5.3, we will need the following lemma, in which $[x_1, y_1, \dots, x_n]$ stands for the set of all infinite continuations (elements of \mathbf{Z}^∞) of x_1, y_1, \dots, x_n .

Lemma 5.4. *For any incomplete data sequence x_1, y_1, \dots, x_n without repeated objects and any $Q \in \mathbf{P}(\mathbf{Z})$,*

$$Q^\infty[x_1, y_1, \dots, x_n] = \sum_{g: \mathbf{X} \rightarrow \mathbf{Y}} Q_g^\infty[x_1, y_1, \dots, x_n] \pi_Q(g). \quad (5.9)$$

Proof. Let \tilde{Q} be the mixture $\sum_{g: \mathbf{X} \rightarrow \mathbf{Y}} Q_g^\infty \pi_Q(g)$. In this proof we will use the notation X_i for the i th random object and Y_i for the i th random label chosen by Reality; x_i and y_i will be the values taken by X_i and Y_i , respectively. (In the rest of the book we mostly use x_i and y_i to serve both goals.) We have, for any $y_n \in \mathbf{Y}$:

$$\begin{aligned} \tilde{Q}[x_1, y_1, \dots, x_n, y_n] &= \prod_{i=1}^n Q_{\mathbf{X}}(x_i) \prod_{i=1}^n \tilde{Q}(Y_i = y_i \\ &\quad | X_1 = x_1, \dots, X_n = x_n, Y_1 = y_1, \dots, Y_{i-1} = y_{i-1}) \\ &= \prod_{i=1}^n Q_{\mathbf{X}}(x_i) \prod_{i=1}^n \tilde{Q}(Y_i = y_i | X_i = x_i) \\ &= \prod_{i=1}^n Q_{\mathbf{X}}(x_i) \prod_{i=1}^n Q_{\mathbf{Y}|\mathbf{X}}(y_i | x_i) = Q[x_1, y_1, \dots, x_n, y_n] \end{aligned}$$

(of course, the second equality is true only because all objects x_1, \dots, x_n are different). \square

The proof shows that, if there are repeated objects in ω , we still have an inequality between the two sides of (5.9): “ \leq ” if the same object is always labeled in the same way in ω , and “ \geq ” otherwise (the latter is obvious, since in this case the right-hand side of (5.9) is zero).

It is easy to derive the statement of Proposition 5.3 from Lemma 5.4.

Proof of Proposition 5.3. Define

$$T_Q(x_1, y_1, \dots, x_n) := \begin{cases} \max_{g: \mathbf{X} \rightarrow \mathbf{Y}} t_{Q_g}(x_1, y_1, \dots, x_n) & \text{if all } x_1, \dots, x_n \text{ are different} \\ 1 & \text{otherwise.} \end{cases}$$

This is an incomplete statistical test since, by Lemma 5.4,

$$\begin{aligned} Q^\infty \{(x_1, y_1, x_2, y_2, \dots) : T_Q(x_1, y_1, \dots, x_n) \leq \epsilon\} \\ = \sum_{g: \mathbf{X} \rightarrow \mathbf{Y}} Q_g^\infty \{(x_1, y_1, x_2, y_2, \dots) : T_Q(x_1, y_1, \dots, x_n) \leq \epsilon\} \pi_Q(g) \\ \leq \sum_{g: \mathbf{X} \rightarrow \mathbf{Y}} Q_g^\infty \{(x_1, y_1, x_2, y_2, \dots) : t_{Q_g}(x_1, y_1, \dots, x_n) \leq \epsilon\} \pi_Q(g) \leq \epsilon. \end{aligned}$$

It is obvious that T will satisfy the requirement of Proposition 5.3. \square

The proof

It remains to derive Theorem 5.2 from Proposition 5.3. Let Γ be a weakly valid probability estimator. Define an incomplete statistical test t by (5.6). Let T be an incomplete statistical test whose existence is guaranteed by Proposition 5.3. Define a probability estimator $\tilde{\Gamma}$ as Γ in (5.7) with t replaced by T .

Let x_1, y_1, \dots, x_n be an incomplete data sequence. If the antecedent of (5.5) holds, t will reject at level ϵ all Q for which $Q_{\mathbf{Y}|\mathbf{X}}(\cdot | x_n)$ is degenerate. By (5.8), T will reject all Q at level ϵ . By the definition of $\tilde{\Gamma}$, $\tilde{\Gamma}^\epsilon(x_1, y_1, \dots, y_n)$ will be empty.

Remark We can see that the problem of the possibility of learning the conditional probabilities for y_n given x_n is simple in two cases: if a significant number of x_1, \dots, x_{n-1} coincide with x_n , we can estimate the conditional probabilities from the available statistics; if all objects x_1, \dots, x_n are different, the task is infeasible, unless the probabilities are degenerate. It would be interesting to study intermediate cases (such as: none of x_1, \dots, x_{n-1} coincides with x_n but there are repetitions among them).

5.4 Bibliographical remarks and addenda

Theorem 5.2 (in the binary case, i.e., Proposition 5.1) was first proved in Nouretdinov et al. 2001b. The original statement of this theorem was given in terms of algorithmic randomness, but later it was strengthened by restating it in more traditional terms (as mentioned in Chap. 2, this is a typical development). The algorithmic result and the idea of its proof were first mentioned in Vovk et al. 1999.

Our proof of Theorem 5.2 was based on the notion of a statistical test; this is, of course, a standard notion of statistics, but in the form used here it was first introduced, perhaps, by Per Martin-Löf (1966) in his version of Kolmogorov's theory of algorithmic randomness.

Density estimation, regression estimation, and regression with deterministic objects

Vapnik (1995, 1998) lists pattern recognition (called classification in this book), regression estimation, and density estimation as the three main learning problems.

The goal of Nouretdinov et al. 2001b was to study the feasibility of these problems under unconstrained randomness.

In the version of the problem of density estimation relevant to the topic of this book (“conditional density estimation”), we start from a measure μ on the label space \mathbf{Y} and our goal is, given a new object x_n , to estimate the density (which is assumed to exist) of the probability distribution of its label y_n w.r. to the measure μ . The standard case is where \mathbf{Y} is a Euclidean space, but Proposition 5.1 shows that this problem is infeasible already in the simplest case $\mathbf{Y} = \{0, 1\}$. Most of the existing literature deals with the case where the objects are absent. For a recent exposition of the theory see Devroye and Lugosi 2001; Vapnik (1998) constructs a version of SVM for density estimation.

There are several possible understandings of the term “regression”. In §2.3 we constructed an efficient confidence predictor for regression under unconstrained randomness. There are, however, two popular understandings (one of them setting a different goal and the other making a different assumption about Reality) that are not feasible for diverse data sets under unconstrained randomness.

One understanding is “regression estimation”: we assume that the examples are generated independently from some probability distribution Q on \mathbf{Z} and our goal is, given (x_i, y_i) , $i = 1, \dots, n - 1$, and x_n , to estimate the conditional expectation of y_n given x_n . This problem is infeasible already in the simple case $\mathbf{Y} = \{0, 1\}$, since it is then coincides with that of probability estimation.

Another understanding is “regression with deterministic objects”. A classic textbook (Cramér 1946) clearly describes two different assumptions that can be made in regression:

- in one approach (Cramér 1946, Chap. 23) it is assumed that the examples (x_i, y_i) are generated by some power distribution;
- in the other approach (Cramér 1946, Chap. 37) it is assumed that the objects x_i are generated by an unknown mechanism (for example, are chosen arbitrarily by the experimenter) and only the labels y_i are generated stochastically; namely, it is assumed that for every $x \in \mathbf{X}$ there is a probability distribution $Q(x)$ on \mathbf{Y} such that y_i is generated from $Q(x_i)$ (formally, Q is required to be a Markov kernel from \mathbf{X} to \mathbf{Y}).

The first model (essentially the model used in Chap. 2) is the combination of the second model (with no assumptions about x_i) and the assumption that x_i are generated by a power distribution. It is easy to see that regression in the second sense is infeasible under unconstrained randomness for diverse data sets. Indeed, if we have a data sequence $(x_1, y_1), \dots, (x_n, y_n)$ such that all x_i are different, it will be perfectly typical (e.g., algorithmically random) w.r. to any Markov kernel Q such that $Q(x_i)$ is concentrated at y_i , $i = 1, \dots, n$; therefore, nontrivial prediction of y_n given x_1, y_1, \dots, x_n is not possible. Of course, the situation changes if additional assumptions are imposed on the Markov kernel Q : cf. the discussion of the Gauss linear model in Chap. 8.

Universal probabilistic predictors

In Chap. 1 we mentioned Stone’s (1977) result about the existence of a universally consistent probabilistic predictor. This result has been improved and extended in

different directions, and its version was used in Chap. 3 (Lemma 3.9 on p. 83). Intuitively, the fact that conditional probabilities can be estimated to any accuracy in the limit, under unconstrained randomness and without assuming precise repetitions, appears to be in some conflict with this chapter's results. From the purely mathematical point of view it suffices to say, as we did in Chap. 3, that convergence in Stone's theorem and its extensions is not uniform. For details, see Devroye et al. 1996 (Chap. 7); we will only illustrate this with a simple example.

Suppose x_n are generated from the uniform distribution on $[0, 1]$ and $y_n = f(x_n)$, where $f : [0, 1] \rightarrow \{0, 1\}$ is a Borel function; we want our inferences to hold without any further assumption about f . Asymptotically, predicting with the majority of the K_n nearest neighbors of x_n (with $K_n \rightarrow \infty$ and $K_n = o(n)$) we will ensure that the frequency of errors ($Q_{Y|X}(1 | x_n) \neq f(x_n)$, in our usual notation) tends to zero (Lemma 3.9 above). Before infinity, however, we cannot say anything at all about the closeness of our predictions to the true conditional probabilities. For any given value of n , no matter how large, there exists an f agreeing with the available data (i.e., such that $f(x_i) = y_i$, $i = 1, \dots, n - 1$) for which $f(x_n) = 0$ and there exists an f agreeing with the available data for which $f(x_n) = 1$.

At a more philosophical and controversial level, it might even be argued that the asymptotic results about universal probabilistic prediction (and so, by implication, some of our results in Chaps. 3 and 6) are devoid of empirical meaning. Let $\mathbf{Y} = \{0, 1\}$. If the object space \mathbf{X} is fixed and finite, the objects in the data sequence z_1, z_2, \dots will eventually start to repeat, no matter how big \mathbf{X} is, and we will then be able to estimate $Q(1 | x_n)$. Kolmogorov's axioms of probability include the axiom of continuity (A.1) (equivalent, in the presence of the other axioms, to σ -additivity). According to Kolmogorov (1933a, p. 15 of the English translation), it is almost impossible to elucidate the empirical meaning of this axiom and its acceptance is an arbitrary, although expedient, choice. (For further details, see Shafer and Vovk 2003.) It appears that the main effect of the acceptance of this axiom was to make infinite probability spaces (the subject of Chap. II of Kolmogorov 1933a) similar to finite probability spaces (the subject of Kolmogorov's Chap. I). Fixing \mathbf{X} (maybe infinite but with the probability distribution generating examples satisfying the axiom of continuity (A.1)) and letting $n \rightarrow \infty$ might be just an embellished version of fixing a finite \mathbf{X} and letting $n \rightarrow \infty$. More relevant, from the empirical point of view, asymptotic results would consider a variable object space \mathbf{X} ; e.g., we could consider a triangular array $x_1, y_1, \dots, x_n, y_n$ where $x_n \in \mathbf{X}_n$, $y_n \in \mathbf{Y}$, and \mathbf{X}_n is an increasing sequence of Borel spaces.

Algorithmic randomness perspective

Theorem 5.2 shows that probabilistic prediction is infeasible in the sense that the conditional probability for the label cannot be estimated, under the stated conditions. A related question is: can we find a conditional probability which is as good as the true conditional probability? The former can be far from the latter, but still be as good in explaining the data. The following result by Ilia Nouretdinov formalizes this question using the algorithmic notion of randomness (Martin-Löf 1966) and shows that the answer is “no”, even if the labels are binary and the objects are absent. Remember that the notion of algorithmic randomness has nothing to do with the assumption of randomness (see the remark on p. 49).

Theorem 5.5. *For any computable probability distribution P on $\{0, 1\}^\infty$ there exist a computable Bernoulli distribution \mathbf{B}_θ on $\{0, 1\}$ and an infinite sequence in $\{0, 1\}^\infty$ which is algorithmically random w.r. to \mathbf{B}_θ^∞ but not algorithmically random w.r. to P .*

The probability distribution P in this theorem is the suggested way of finding a good conditional probability (since a conditional probability can be found for any data sequence, these conditional probabilities can be put together to form a probability distribution). The theorem asserts that there exists a power distribution (even a computable one) and an infinite sequence which can be produced by this power distribution but for which P is not a good explanation.

Proof. If $\omega = (y_1, y_2, \dots) \in \{0, 1\}^\infty$ is an infinite binary sequence, we set

$$\theta_N(\omega) := \frac{y_1 + \dots + y_N}{N}.$$

Let \mathbf{B}_θ be the Bernoulli distribution on $\{0, 1\}$ corresponding to a parameter $\theta \in [0, 1]$. For each $n = 1, 2, \dots, N(n)$ is defined constructively as any number N such that, for any $\theta \in [0, 1]$,

$$\mathbf{B}_\theta^\infty \{ \omega : |\theta_N(\omega) - \theta| > 2 \times 10^{-n-1} \} \leq 2^{-n-1}.$$

Set

$$[a_0, b_0] := [0, 1]$$

and

$$U_0 := \{0, 1\}^\infty.$$

Define inductively, for $n = 1, 2, \dots$,

$$[a_n, b_n] := [a_{n-1} + 2 \times 10^{-n}, a_{n-1} + 3 \times 10^{-n}] \quad (5.10)$$

or

$$[a_n, b_n] := [a_{n-1} + 7 \times 10^{-n}, a_{n-1} + 8 \times 10^{-n}] \quad (5.11)$$

and then

$$U_n := \{ \omega \in U_{n-1} : \theta_{N(n)}(\omega) \in [a_n, b_n] \},$$

where the choice between (5.10) and (5.11) is done effectively and so that $P(U_n)$ decreases significantly (say, $P(U_n) \leq \frac{2}{3}P(U_{n-1})$). Finally, set

$$\theta := \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n$$

and

$$U := \cap_n U_n.$$

The statement of the theorem follows from the following properties of this construction:

1. U is a constructively null set ($P(U) = 0$ since $P(U_n) \leq \frac{2}{3}P(U_{n-1})$ for all n) and θ is computable.
2. U contains some algorithmically random sequences w.r. to \mathbf{B}_θ^∞ .

Only the last property requires a separate proof.

Using $\mathbf{B}_\theta^\infty(U_0) = 1$ and the fact that

$$a_n + 2 \times 10^{-n-1} \leq a_{n+1} \leq \theta \leq b_{n+1} \leq b_n - 2 \times 10^{-n-1},$$

we obtain

$$\begin{aligned}\mathbf{B}_\theta^\infty(U_{n-1} \setminus U_n) &\leq \mathbf{B}_\theta^\infty\{\omega : \theta_{N(n)}(\omega) \notin [a_n, b_n]\} \\ &\leq \mathbf{B}_\theta^\infty\{\omega : |\theta_{N(n)}(\omega) - \theta| > 2 \times 10^{-n-1}\} \leq 2^{-n-1};\end{aligned}$$

therefore, $\mathbf{B}_\theta^\infty(U) \geq 1/2$. □

Probabilistic prediction II: Venn predictors

We saw in Chap. 1 that there are different types of algorithms for learning under randomness, where examples (objects with labels) are drawn one by one from an unknown probability distribution. Appropriate criteria for validity and efficiency may vary from type to type. In Chap. 2, we developed criteria for the validity and efficiency of confidence predictors, and we developed a class of confidence predictors, conformal predictors, that satisfy the criterion for validity while varying in their efficiency. In this chapter, we study algorithms of a new type, multiprobability predictors. We develop a criterion of validity for multiprobability predictors, and we introduce and study a class of multiprobability predictors, Venn predictors, that satisfy the criterion for validity when the label space is finite. Venn predictors also vary in their efficiency, but we will see that there are arguments, both theoretical and empirical, for the efficiency of some simple Venn predictors; therefore, the impossibility of estimation of non-extreme probabilities in a diverse random environment (Theorem 5.2) does not prevent us from producing probabilities, perhaps quite different from the “true” ones, which perform well in important respects.

Venn predictors use a familiar idea: divide the old objects into categories, somehow classify the new object into one of the categories, and then use the frequencies of labels in the chosen category as probabilities for the new object’s label. We innovate only in a couple of details:

- We divide examples rather than objects into categories. When we compute the frequencies of labels in the category containing the new example, we include the new example along with the old examples already in that category. Since at the time of prediction we do not yet know the new object’s label, we compute these frequencies several times, once for each label the new object might have. (This is analogous to the way we treat the new object when we use a nonconformity measure to define a conformal predictor.)

- We interpret each set of frequencies as a probability distribution for the new object's unknown label. Thus we announce several probability distributions for the new label rather than a single one. Fortunately, once the number of old examples in each category is large, these different probability distributions will be practically identical.

The task of valid multiprobability prediction, which is achieved by Venn predictors, can be contrasted with the more demanding task of valid probabilistic prediction. In valid probabilistic prediction, we announce a single probability distribution for each new label y_n , $n = 1, 2, \dots$, and these probability distributions are supposed to perform well against statistical tests based on the subsequent observation of the labels. We cannot expect to achieve valid probabilistic prediction under unconstrained randomness (even in the binary case, as was shown in Theorem 5.5). Our criterion for validity for multiprobability prediction lowers the bar in two respects. First, we gain some wiggle room when we are allowed to announce several probability distributions; we get by whenever one of them is acceptable. Second, instead of being exposed to arbitrary tests, we are exposed only to tests of calibration. Tests of calibration check only whether probabilities are matched by observed frequencies; for example, a particular label should occur in about 25% of the instances in which we give it a probability near 0.25.

The first section of this chapter, §6.1, studies on-line probabilistic prediction in some depth. Here we discuss how on-line probabilistic predictions can be tested using supermartingales and how to identify the supermartingales that test calibration. We also state an important impossibility result: under unconstrained randomness, there is no strategy for probabilistic prediction that will perform well even against tests of calibration. This impossibility result motivates our replacing single probabilities with multiple probabilities.

In §6.2, we formulate our criterion of validity for multiprobability prediction. This involves a straightforward extension of the concept of supermartingale testing from single probability distributions to multiple probability distributions.

In §6.3, we define precisely the class of Venn predictors and show that these predictors do provide multiprobability predictions that are valid according to our criterion. Our discussion in this section includes empirical results on the USPS data set; we find that the Venn predictor based on the nearest neighbor idea performs reasonably well on this data set. At the end of §6.3 we look more carefully at the contrast between Venn predictors, which provide numbers that have some but not all of the properties of probabilities, and conformal predictors, which provide numbers that are even less like probabilities.

In §6.4, we give an asymptotic result: there exists a Venn predictor that asymptotically approaches the true conditional probabilities. Unfortunately, this result is impractical in the same sense as the other strong asymptotic results we discuss in this book (see Chap. 5).

In Chap. 9, we will extend the concepts introduced in this chapter to on-line compression models.

6.1 On-line probabilistic prediction

Probabilistic prediction, or probability forecasting, as it is often called, has been studied for several decades (see the review by Dawid 1986). The existing literature often emphasizes asymptotic theory, but as we show in this section, there are reasonable ways to test finite sequences of probabilistic predictions, and in particular there are reasonable ways to test them for calibration.

The framework we use for probabilistic prediction in this section is very close to the framework we introduced in Chap. 2. We observe a sequence of examples, each example z_n consisting of an object x_n and its label y_n . At each trial, we first see the object and want to predict the label. The object is drawn from a measurable space \mathbf{X} , and the label is drawn from a measurable space \mathbf{Y} , so that $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$. But there are three points of difference from Chap. 2:

1. As in Chap. 3, we now deal with the problem of classification, assuming that the label space \mathbf{Y} is finite. (We carry this assumption throughout this chapter.)
2. Instead of imagining that the sequence of examples z_1, z_2, \dots continues indefinitely, we now assume that we will observe only a finite number of examples, say z_1, \dots, z_N , and that we know the number N , called the *horizon*, in advance. (We carry this assumption through most of the chapter, dropping it only in §6.4.)
3. We do not assume that the examples (z_1, \dots, z_N) are drawn from an exchangeable probability distribution. (We will come back to the assumption that the examples are drawn from an exchangeable, or even power, probability distribution, but this assumption is out of place in a general treatment of probabilistic prediction.)

We assume that the probability distribution P on \mathbf{Z}^N comes supplied with regular conditional probabilities (see §A.4; they exist because \mathbf{Y} is finite), so that there is no ambiguity when we speak of the regular conditional distribution of y_n given x_1, y_1, \dots, x_n ; for $y \in \mathbf{Y}$, we will write $P(y | x_1, y_1, \dots, x_n)$ for the conditional P -probability that $y_n = y$ given x_1, y_1, \dots, x_n (in particular, $P(y_n | x_1, y_1, \dots, x_n)$ will be the conditional probability of the realized label y_n given x_1, y_1, \dots, x_n).

This is a long section, with a number of subsections. In the first subsection (p. 146), we formulate the task of probabilistic prediction in terms of a game protocol. In the following subsection (p. 147), we use a simplified version of this protocol, where \mathbf{Y} has only two elements, to discuss informally how probabilistic predictions can be tested and in particular how their calibration can be tested. Then we look at two ways of formalizing the notion of a test:

we can check whether an event of small probability occurs (p. 148), or we can check whether a nonnegative martingale gets much larger than its initial value (p. 150). In both cases, we consider the test to be a test of calibration if it does not depend on the order of the examples. Finally, in the last subsection (p. 154), we state our negative result: no strategy for probabilistic prediction performs well against tests of calibration.

There are no results in this section that will be used in the rest of the chapter; but what we learn here motivates our definitions in §6.2. The fact that we cannot perform well against tests of calibration when we predict with single probability distributions motivates the consideration of a set-up where we predict with multiple probability distributions, and the martingale concepts developed here motivate the corresponding concepts for multiprobability prediction. In addition, the definitions and notation for game martingales that we introduce in this section (see p. 150) are taken for granted in §6.2.

The on-line protocol for probabilistic prediction

Probabilistic prediction means giving probabilities for outcomes before the outcomes become known. In our on-line setting, where we are observing $x_1, y_1, \dots, x_N, y_N$ in sequence and we have just observed x_n , this means giving a probability distribution, say p_n , for y_n in light of the previous information, x_1, y_1, \dots, x_n . This is the task of Predictor in the following protocol, where we write $\mathbf{P}(\mathbf{Y})$ for the set of all probability distributions on \mathbf{Y} :

```

PROBABILISTIC PREDICTION
FOR  $n = 1, 2, \dots, N$ :
    Reality announces  $x_n \in \mathbf{X}$ ;
    Predictor announces  $p_n \in \mathbf{P}(\mathbf{Y})$ ;
    Reality announces  $y_n \in \mathbf{Y}$ 
END FOR.

```

If Predictor knew the probability distribution P on \mathbf{Z}^N from which the sequence $x_1, y_1, \dots, x_N, y_N$ is drawn, then he would, of course, set his p_n equal to P 's conditional probability for y_n given x_1, y_1, \dots, x_n . This would not guarantee that Predictor's p_n would pass a given test; the y_n might by chance come out in such a way that the p_n would be rejected. But this is not too likely, and we can still say that Predictor is performing perfectly when he announces the true conditional probabilities.

Because Predictor does not know P , he cannot perform perfectly. We can nevertheless hold perfect performance up as his goal and test whether the observed sequence $x_1, y_1, \dots, x_N, y_N$ is consistent with the hypothesis that it is drawn from a distribution P that has Predictor's p_1, \dots, p_N as its appropriate conditional probabilities. After all, the ability to withstand all kinds of tests is the only possible empirical meaning for the assumption that a sequence is drawn from a probability distribution with certain properties.

At the end of this section, we will turn our attention to strategies for Predictor – rules that tell Predictor how to choose p_n in light of the previous moves by Reality, x_1, y_1, \dots, x_n . We would like to have Predictor’s strategy whose probabilistic predictions stand up to tests of calibration as well as possible.

An informal look at testing calibration

Consider for a moment the case where \mathbf{Y} has only two elements, say $\mathbf{Y} = \{0, 1\}$. In this case, we can write p_n for the probability Predictor announces for the event $y_n = 1$, and our protocol takes this form:

```
BINARY PROBABILISTIC PREDICTION
FOR  $n = 1, 2, \dots, N$ :
    Reality announces  $x_n \in \mathbf{X}$ ;
    Predictor announces  $p_n \in [0, 1]$ ;
    Reality announces  $y_n \in \{0, 1\}$ 
END FOR.
```

This simplified protocol is useful for an informal discussion of some standard ideas about testing the calibration of probabilistic predictions.

One obvious question is whether Predictor’s p_n tend to be too high or too low overall. We can test this by comparing their overall average,

$$\bar{p}_N := \frac{1}{N} \sum_{n=1}^N p_n ,$$

with the overall frequency of 1s among the y_n ,

$$\bar{y}_N := \frac{1}{N} \sum_{n=1}^N y_n .$$

If

$$\bar{y}_N \approx \bar{p}_N , \tag{6.1}$$

then we may say that the p_n are “unbiased on average”. If the approximate inequality (6.1) is violated – if \bar{y}_N and \bar{p}_N are too different – then we may reject the p_n on the grounds that they are biased on average.

A more refined test of calibration would look at the subset of n for which p_n is close to a given value p^* , and compare the frequency of $y_n = 1$ in this subset, say $\bar{y}_N(p^*)$, with p^* . If

$$\bar{y}_N(p^*) \approx p^* \text{ for all } p^* , \tag{6.2}$$

then the p_n can be considered “well calibrated”. Probabilistic predictions that pass this test at least get the frequencies right; in this sense they tell the truth.

It is easy to see, however, that good calibration is not enough to make probabilistic predictions useful. A popular example for demonstrating this point assumes that it is known in advance that the labels will follow the pattern

$$y_n = \begin{cases} 1 & \text{if } n \text{ is odd} \\ 0 & \text{otherwise.} \end{cases} \quad (6.3)$$

In this example, Predictor can achieve excellent calibration by always setting $p_n = 0.5$. But the probabilistic predictions $p_1 = 1, p_2 = 0, \dots$ are preferable, because they predict y_n exactly with certainty.

What we need in addition to good calibration is sometimes called “high resolution”. This term suggests a procedure like the one we will use for our Venn predictor in §6.3: Predictor sorts the objects into categories and uses the frequencies in the category where a new object x_n falls as his probabilistic prediction p_n . The more numerous the categories – the more information Predictor takes into account in creating the categories – the greater the resolution and the more useful the probabilistic prediction should be. In the example given by (6.3), Predictor should use two categories and predict y_n perfectly rather than using one category and always giving the probability 0.5. The shortcoming of the term “resolution” is that this procedure may be too special. It is not always better to have more categories, and we do not want to rule out Predictor’s using some completely different kind of algorithm for calculating his probabilities. So rather than speak of resolution, we will speak of efficiency. Probabilistic predictions are efficient when they are as informative – as close to zero or one – as possible.

Testing using events of small probability

In general, a statistical test of a hypothesis is defined by identifying a set that has a small probability if the hypothesis is true. According to *Cournot’s principle* (Shafer and Vovk 2003), the hypothesis may be rejected if the event of small probability happens.

Let us now consider how this idea applies to our general probabilistic prediction protocol, where \mathbf{Y} is finite but not necessarily binary. To begin, we set

$$\Pi := (\mathbf{P}(\mathbf{Y}) \times \mathbf{Y})^N. \quad (6.4)$$

This is the space of all sequences of play $p_1, y_1, \dots, p_N, y_N$ with all the objects omitted (since they are not involved¹ in the idea of calibration). We call Π the *game space*, and we call any measurable subset of Π a *game event*. Given a game event E and a probability distribution R on \mathbf{Z}^N with specified regular conditional probabilities $R(\cdot | z_1, \dots, z_{n-1}, x_n) \in \mathbf{P}(\mathbf{Y})$ for different possible values of y_n , we set

¹It might be interesting to consider “conditional” notions of calibration (within natural categories of examples), but we only consider the simplest case.

$$\tilde{R}(E) := R \left\{ (x_1, y_1, \dots, x_N, y_N) \in \mathbf{Z}^N : \right. \\ \left. (R(\cdot | x_1), y_1, R(\cdot | z_1, x_2), y_2, \dots, R(\cdot | z_1, \dots, z_{N-1}, x_N), y_N) \in E \right\}.$$

This is probability that E will happen if the sequence of examples is drawn from R and Predictor uses R 's specified regular conditional distributions as his strategy. We define the *upper probability* $\bar{\mathbb{P}}(E)$ of a game event E by

$$\bar{\mathbb{P}}(E) := \sup_R \tilde{R}(E), \quad (6.5)$$

where the supremum is over all probability distributions with regular conditional probabilities – i.e., over all ways of specifying the probability distribution R and all ways of specifying regular conditional probabilities for it.

The upper probability $\bar{\mathbb{P}}(E)$ is the highest probability E can have if the p_n are appropriate conditional probabilities for whatever probability distribution governs the x_n and y_n . If $\bar{\mathbb{P}}(E)$ is small – say less than some specified small probability δ – and E happens, we invoke Cournot's principle and reject at level δ the hypothesis that the p_n are appropriate conditional probabilities. In this case we may say that E is a *level δ statistical test*; we will sometimes call δ the *significance level* of the test.

Example 6.1. To illustrate these definitions, consider the problem of testing average unbiasedness for the binary case, as defined by (6.1). Here we plan to reject Predictor's p_1, \dots, p_N if $|\bar{y}_N - \bar{p}_N|$ is too large. There are many ways of deciding how large is too large: we can use Chebyshev's inequality, the central limit theorem, large deviation inequalities, etc. As an example, we consider a particular large deviation inequality, Hoeffding's inequality (see §A.7). This inequality implies that for any $\epsilon > 0$,

$$\mathbb{P}\{|\bar{y}_N - \bar{p}_N| \geq \epsilon\} \leq 2 \exp(-2\epsilon^2 N), \quad (6.6)$$

where p_n is the conditional probability that $y_n = 1$ given x_1, y_1, \dots, x_n . It follows that

$$\mathbb{P}\left\{|\bar{y}_N - \bar{p}_N| \geq \sqrt{\frac{\ln \frac{2}{\delta}}{2N}}\right\} \leq \delta$$

for any $\delta > 0$ whenever p_n is the conditional probability that $y_n = 1$ given x_1, y_1, \dots, x_n . Hence the game event

$$E := \left\{ (p_1, y_1, \dots, p_N, y_N) : |\bar{y}_N - \bar{p}_N| \geq \sqrt{\frac{\ln \frac{2}{\delta}}{2N}} \right\} \quad (6.7)$$

has upper probability $\bar{\mathbb{P}}(E)$ less than or equal to δ . When E happens, we can invoke Cournot's principle and reject the p_n at level δ .

Calibration events

We call a game event E a *calibration event* if it is invariant with respect to permutations – i.e., if

$$(p_1, y_1, \dots, p_N, y_N) \in E \implies (p_{\pi(1)}, y_{\pi(1)}, \dots, p_{\pi(N)}, y_{\pi(N)}) \in E$$

for any sequence $(p_1, y_1, \dots, p_N, y_N) \in \Pi$ and any permutation π of the set $\{1, \dots, N\}$. Intuitively, calibration events are the events that can be used to test calibration.

The game event (6.7) is a calibration event in the binary protocol; more generally, any natural formalization of (6.1) and (6.2) will give a calibration event.

Testing using nonnegative supermartingales

As we will see later in this chapter (Example 6.4) and in the following chapter, on-line tests can often be described more conveniently in terms of nonnegative martingales or nonnegative supermartingales than in terms of events of small probability. A very brief introduction to the standard theory of martingales is given in §A.6, but in this chapter we will also need the following less standard definitions:

- A *game martingale* is a measurable function G on the sequences of the form $p_1, y_1, \dots, p_n, y_n$, where $n = 0, \dots, N$, $p_i \in \mathbf{P}(\mathbf{Y})$, and $y_i \in \mathbf{Y}$, that satisfies

$$G(p_1, y_1, \dots, p_{n-1}, y_{n-1}) = \int_{\mathbf{Y}} G(p_1, y_1, \dots, p_n, y) p_n(dy) \quad (6.8)$$

for all p_1, y_1, \dots, p_n , $n = 1, \dots, N$.

- A *game supermartingale* is a measurable function G on the same domain that satisfies

$$G(p_1, y_1, \dots, p_{n-1}, y_{n-1}) \geq \int_{\mathbf{Y}} G(p_1, y_1, \dots, p_n, y) p_n(dy) \quad (6.9)$$

for all p_1, y_1, \dots, p_n , $n = 1, \dots, N$.

- A *calibration martingale* (resp., *calibration supermartingale*) is a nonnegative game martingale (resp., nonnegative game supermartingale) whose final value is invariant under permutations:

$$G(p_1, y_1, \dots, p_N, y_N) = G(p_{\pi(1)}, y_{\pi(1)}, \dots, p_{\pi(N)}, y_{\pi(N)})$$

for any $p_1, y_1, \dots, p_N, y_N$ and any permutation π of $\{1, \dots, N\}$.

We can define the upper probability of a game event E in terms of game martingales or game supermartingales:

$$\begin{aligned}\bar{\mathbb{P}}(E) := \inf \{ G(\square) : & G(p_1, y_1, \dots, p_N, y_N) \geq 1, \\ & \forall (p_1, y_1, \dots, p_N, y_N) \in E \} ,\end{aligned}\quad (6.10)$$

where \square is the empty sequence and G ranges over all nonnegative game martingales, or, equivalently, where G ranges over all nonnegative game supermartingales. In the case where \mathbf{X} is finite and p_n are constrained to a finite, as dense as we wish, subset of $\mathbf{P}(\mathbf{Y})$, definition (6.10) is equivalent to the definition we gave in the preceding subsection, (6.5) (see Lemma 6.8 on p. 164). When the difference between the two definitions is essential, we will use $\bar{\mathbb{P}}^{\text{meas}}$ to denote (6.5) (the “measure-theoretic definition”) and $\bar{\mathbb{P}}^{\text{game}}$ to denote (6.10) (the “game-theoretic definition”). Doob’s inequality (p. 285) shows that it is always true that $\bar{\mathbb{P}}^{\text{meas}} \leq \bar{\mathbb{P}}^{\text{game}}$.

The next proposition shows that, for calibration events, $\bar{\mathbb{P}}^{\text{game}}(E)$ can be equivalently defined by letting G range over the calibration supermartingales in (6.10).

Proposition 6.2. *If E is a calibration event, $\bar{\mathbb{P}}^{\text{game}}(E)$ equals the right-hand side of (6.10) where G ranges over the calibration supermartingales.*

Proof. Let E be a calibration event and a game supermartingale G satisfies

$$G(p_1, y_1, \dots, p_N, y_N) \geq 1, \quad \forall (p_1, y_1, \dots, p_N, y_N) \in E .$$

Set

$$\begin{aligned}G^*(p_1, y_1, \dots, p_n, y_n) := \inf_{\pi} G(p_{\pi(1)}, y_{\pi(1)}, \dots, p_{\pi(n)}, y_{\pi(n)}), \\ n = 0, 1, \dots, N ,\end{aligned}\quad (6.11)$$

π ranging over the permutations of $\{1, \dots, n\}$. The measurability of G^* follows from there being only finitely many π in (6.11); it is clear that

$$G^*(p_1, y_1, \dots, p_N, y_N) \geq 1, \quad \forall (p_1, y_1, \dots, p_N, y_N) \in E ;$$

therefore, it only remains to prove that

$$G^*(p_1, y_1, \dots, p_{n-1}, y_{n-1}) \geq \int_{\mathbf{Y}} G^*(p_1, y_1, \dots, p_{n-1}, y_{n-1}, p_n, y) p_n(dy) .\quad (6.12)$$

Let π be a permutation of $\{1, \dots, n-1\}$ at which the infimum in the definition of the left-hand side of (6.12) is achieved. We then have

$$\begin{aligned}G^*(p_1, y_1, \dots, p_{n-1}, y_{n-1}) &= G(p_{\pi(1)}, y_{\pi(1)}, \dots, p_{\pi(n-1)}, y_{\pi(n-1)}) \\ &\geq \int_{\mathbf{Y}} G(p_{\pi(1)}, y_{\pi(1)}, \dots, p_{\pi(n-1)}, y_{\pi(n-1)}, p_n, y) p_n(dy) \\ &\geq \int_{\mathbf{Y}} G^*(p_1, y_1, \dots, p_{n-1}, y_{n-1}, p_n, y) p_n(dy) .\end{aligned}$$

□

We may use game martingales or, more generally, game supermartingales to test Predictor's p_n : we reject the p_n at level $\delta \in (0, 1)$ if a nonnegative game supermartingale that starts at δ becomes 1 or more at the end of the protocol. Equivalently, we reject the p_n at level δ if a nonnegative game supermartingale that starts at 1 becomes $1/\delta$ or more at the end; such testing procedures may be called *level δ martingale tests*. In principle, it is not necessary to fix a threshold $1/\delta$ in advance; we can just interpret the value attained by a nonnegative game supermartingale starting at 1 as measuring the strength of evidence against p_n . If we want to test for calibration, we should use a calibration supermartingale.

The concepts of game martingale and game supermartingale have a natural interpretation in terms of betting (analogous to the usual betting interpretation of the concepts of martingale and supermartingale in standard probability theory). To see this, imagine a player, say Gambler, who bets on y_n at the probabilities given by p_n right after Predictor announces p_n . If Gambler is allowed to bet at odds that are fair according to the probability distribution p_n , then his payoff will have expected value zero relative to p_n :

FOR $n = 1, 2, \dots, N$:

Predictor announces $p_n \in \mathbf{P}(\mathbf{Y})$;

Gambler announces $f_n : \mathbf{Y} \rightarrow \mathbb{R}$ with $\int_{\mathbf{Y}} f_n(y) p_n(dy) = 0$;

Reality announces $y_n \in \mathbf{Y}$

END FOR.

If we write $G(p_1, y_1, \dots, p_n, y_n)$ for Gambler's capital at the end of the n th trial, then

$$G(p_1, y_1, \dots, p_n, y_n) = G(p_1, y_1, \dots, p_{n-1}, y_{n-1}) + f_n(y_n), \quad (6.13)$$

and hence (6.8) holds. If Gambler is allowed to throw money away or to make bets that are unfavorable to himself rather than fair, then we obtain instead (6.9).

We conclude this subsection with two examples of testing calibration using calibration supermartingales.

Example 6.3. To test (6.1) using martingales, let us again turn to Hoeffding's inequality; this time, however, we will need its proof rather than statement. According to (A.12) (on p. 288) and its derivation,

$$\exp\left(\pm\epsilon(\bar{y}_n - \bar{p}_n)n - \frac{\epsilon^2 n}{8}\right) \quad (6.14)$$

are both game supermartingales, for any $\epsilon > 0$, and their sum is evidently nonnegative and invariant under permutations of the examples. Any multiple of their sum is therefore a calibration supermartingale. It follows that rejecting when

$$\exp\left(\epsilon(\bar{y}_N - \bar{p}_N)N - \frac{\epsilon^2 N}{8}\right) + \exp\left(\epsilon(\bar{p}_N - \bar{y}_N)N - \frac{\epsilon^2 N}{8}\right) \geq \frac{2}{\delta}$$

is a level δ martingale test of calibration. If we set $\epsilon := 2\sqrt{2 \ln(2/\delta)/N}$, then this level δ martingale test rejects whenever the level δ statistical test (6.7) rejects.

Example 6.4. The problem of testing calibration using a partition of the trials based on the values of p_n , (6.2), is subtler than the problem of testing overall unbiasedness, (6.1).

An obvious first step is to split the interval $[0, 1]$ into K bins

$$B_k := \left[\frac{k-1}{K}, \frac{k}{K} \right), \quad k = 1, \dots, K-1, \quad B_K := \left[\frac{K-1}{K}, 1 \right] \quad (6.15)$$

of equal width $1/K$ and look at the deviations

$$\left| \frac{1}{N_k} \sum_{n \in \{1, \dots, N\}: p_n \in B_k} (y_n - p_n) \right|, \quad k = 1, \dots, K, \quad (6.16)$$

where $N_k := |\{n = 1, \dots, N : p_n \in B_k\}|$ is the number of p_n that fall in B_k .

If we now try to bound the probabilities for the events (6.16), however, we face the serious difficulty that the inequalities that do this, such as Hoeffding's inequality (see (A.9) and (A.10) on p. 287), require that the number of examples N be fixed in advance, whereas the N_k in (6.16) are random. We might choose a threshold T and apply Hoeffding's inequality to the first T examples in the bin B_k if $N_k \geq T$, but it unclear how large T should be. We have no *a priori* estimate of the magnitude of N_k , and it seems likely that however we choose T , either N_k will be much larger than T (in which case we will be using the loose upper bound $2 \exp(-2\epsilon^2 T)$ instead of the desired $2 \exp(-2\epsilon^2 N_k)$) or N_k will fall short of T (in which case we will forfeit this opportunity to test calibration altogether).

Supermartingales are flexible enough to avoid this difficulty. Calibration supermartingales testing (6.2) are easily constructed from the game supermartingales (6.14). The most natural one is perhaps

$$\begin{aligned} & \frac{1}{2K} \left(\sum_{k=1}^K \exp \left(\epsilon \sum_{i \in \{1, \dots, n\}: p_i \in B_k} (y_i - p_i) - \frac{\epsilon^2 N_k^n}{8} \right) \right. \\ & \quad \left. + \sum_{k=1}^K \exp \left(\epsilon \sum_{i \in \{1, \dots, n\}: p_i \in B_k} (p_i - y_i) - \frac{\epsilon^2 N_k^n}{8} \right) \right), \quad (6.17) \end{aligned}$$

where $N_k^n := |\{i = 1, \dots, n : p_i \in B_k\}|$ is the number of the p_i among the first n that fall in the k th bin.

At the intuitive level, a reasonable choice of ϵ would be of the same order of magnitude as the expected maximum deviation in different bins of the average probabilistic prediction from the average label. If, for example, the average

label is greater than the average prediction by $\epsilon/4$ in the k th bin, the final value of the calibration supermartingale (6.17) will be at least

$$\frac{1}{2K} \exp \left(\epsilon \sum_{i \in \{1, \dots, N\}: p_i \in B_k} (y_i - p_i) - \frac{\epsilon^2 N_k}{8} \right) \geq \frac{1}{2K} \exp \left(\frac{\epsilon^2 N_k}{8} \right),$$

which is large when N_k is large enough and K is moderately large. If the best choice of ϵ is not clear, we can always “mix” the calibration supermartingales for different values of ϵ , and instead of (6.17) use, e.g.,

$$\int_0^1 \frac{1}{2K} \left(\sum_{k=1}^K \exp \left(\epsilon \sum_{i \in \{1, \dots, n\}: p_i \in B_k} (y_i - p_i) - \frac{\epsilon^2 N_k^n}{8} \right) + \sum_{k=1}^K \exp \left(\epsilon \sum_{i \in \{1, \dots, n\}: p_i \in B_k} (p_i - y_i) - \frac{\epsilon^2 N_k^n}{8} \right) \right) \mu(d\epsilon)$$

for a suitable probability distribution μ on $[0, 1]$. □

Predictor has no satisfactory strategy

We turn now to the question of how well a strategy for Predictor can perform with respect to calibration. As we warned at the beginning of the chapter, we will give the obvious negative answer: no strategy for Predictor can produce probabilistic predictions that look as well calibrated as the ones Predictor could produce if he knew the probability distribution from which the examples are drawn.

Formally, a strategy for Predictor, or a *probabilistic predictor*, as we are calling it, is a Markov kernel F that assigns a probability distribution

$$F(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \in \mathbf{P}(\mathbf{Y})$$

to each sequence $(x_1, y_1, \dots, x_n) \in \mathbf{Z}^{n-1} \times \mathbf{X}$, for $n = 1, \dots, N$. It tells Predictor to use this probability distribution as his move p_n when Reality has previously made the moves x_1, y_1, \dots, x_n . (We already used the notion of a probabilistic predictor in discussions of Chaps. 1 and 5, but without a formal definition.)

We now assume that the individual examples (x_n, y_n) are drawn independently from some distribution Q on \mathbf{Z} , so that the entire sequence $x_1, y_1, \dots, x_N, y_N$ is drawn from the power distribution Q^N on \mathbf{Z}^N .

If Predictor knew Q , he would use the conditional probabilities $Q_{\mathbf{Y}|\mathbf{X}}(\cdot | x_n)$ given by Q , and he would then have a reasonable expectation that his probabilistic predictions would look well calibrated: for any calibration event E , the probability under Q^N of E happening would not exceed $\bar{P}(E)$. He might be unlucky; we might test his calibration using E , and E might happen,

and then we would proclaim him poorly calibrated. But the chance of this happening would be no greater than advertised by the significance level of our test.

Ideally, we want a probabilistic predictor F that performs equally well even though it is defined without knowledge of the true Q . More precisely, we want F to satisfy this condition: for any probability distribution Q on \mathbf{Z} and any calibration event E , the Q^N -probability that

$$(F(x_1), y_1, \dots, F(x_1, y_1, \dots, x_N), y_N) \in E \quad (6.18)$$

never exceeds $\bar{\mathbb{P}}(E)$. We call a probabilistic predictor F that satisfies this condition *weakly N-calibrated*.

Formally, there are two versions of the notion of a weakly N -calibrated strategy: measure-theoretic, in which $\bar{\mathbb{P}}(E)$ is understood as $\bar{\mathbb{P}}^{\text{meas}}(E)$, and game-theoretic, in which $\bar{\mathbb{P}}(E)$ is understood as $\bar{\mathbb{P}}^{\text{game}}(E)$. Here is our negative result, which holds for both versions:

Theorem 6.5. *No weakly N -calibrated probabilistic predictor exists.*

In order to set the stage for the criterion for validity that we will formulate for multiprobability prediction in the next section, let us restate the game-theoretic definition of weak N -calibration. Because of Theorem 6.5, this discussion will be vacuous from the formal point of view; we believe that it still has a heuristic value, but the reader might wish to skip the rest of this section.

Let \mathcal{F}_n be the σ -algebra generated by the first n examples z_1, \dots, z_n , $n = 1, \dots, N$. We will say “ P -supermartingale” to mean a supermartingale in the probability space $(\mathbf{Z}^N, \mathcal{F}_N, P)$ w.r. to the filtration $\mathcal{F}_1, \dots, \mathcal{F}_N$.

Note that $\bar{\mathbb{P}}^{\text{game}}(E) \leq 1/C$ for a calibration event E means that there exists a calibration supermartingale G that starts at one and reaches C whenever E happens, whereas $Q^N(A) \leq 1/C$ for the event $A \subseteq \mathbf{Z}^N$ defined by (6.18) means (by Ville’s theorem on p. 285) that there exists a nonnegative Q^N -supermartingale S that starts at one and reaches C whenever A happens. (For simplicity, we assume that the infima in (6.9) and (A.7) are attained.) Thus the Q^N -probability of (6.18) is less than or equal to $\bar{\mathbb{P}}^{\text{game}}(E)$ if and only if the existence of such a G implies the existence of such an S . Because this is required for any calibration event E , we get this statement: A probabilistic predictor F is weakly N -calibrated if and only if for any probability distribution Q on \mathbf{Z} , any calibration supermartingale G with $G(\square) = 1$, and any threshold $C > 0$ there exists a nonnegative Q^N -supermartingale $S = (S_0, S_1, \dots, S_N)$ with $S_0 = 1$ such that

$$\begin{aligned} G(F(x_1), y_1, \dots, F(x_1, y_1, \dots, x_N), y_N) &\geq C \\ \implies S_N(x_1, y_1, \dots, x_N, y_N) &\geq C \end{aligned}$$

for all $x_1, y_1, \dots, x_N, y_N$.

Once the definition of weak N -calibration is put in this form, it is natural to strengthen it by requiring that the same S work for all thresholds C . This produces the following definition. A probabilistic predictor F is *N -calibrated* if for any probability distribution Q on \mathbf{Z} and any calibration supermartingale G with $G(\square) = 1$ there exists a Q^N -supermartingale S_0, \dots, S_N with $S_0 = 1$ such that

$$G(F(x_1), y_1, \dots, F(x_1, y_1, \dots, x_N), y_N) \leq S_N(x_1, y_1, \dots, x_N, y_N) \quad (6.19)$$

for all $x_1, y_1, \dots, x_N, y_N$. (The supermartingale S is necessarily nonnegative.) Intuitively, this says that if a sequence $x_1 y_1 \dots x_N y_N$ evidences some degree of F 's miscalibration, then it is untypical of Q^N to the same degree.

As it turns out, we will find it convenient to use a concept that is even a bit stronger than N -calibration: first, we replace power distributions Q^N with exchangeable distributions on \mathbf{Z}^N , and second, we replace calibration supermartingales with nonnegative “reversible” supermartingales, in the sense of the following definition. Let us say that a game martingale or supermartingale G is *reversible* if its final value does not change when the order of labels and probabilistic predictions is reversed:

$$G(p_1, y_1, \dots, p_N, y_N) = G(p_N, y_N, \dots, p_1, y_1)$$

for all $(p_1, y_1, \dots, p_N, y_N) \in \Pi$; we will usually abbreviate “reversible game martingale” and “reversible game supermartingale” by omitting “game”. The requirement that G be nonnegative and reversible is evidently weaker than the requirement that it be nonnegative and invariant under permutations of the examples. So the following definition demands more than N -calibration. A probabilistic predictor F is *strongly N -calibrated* if for any exchangeable probability distribution P on \mathbf{Z}^N and any nonnegative reversible supermartingale G with $G(\square) = 1$ there exists a P -supermartingale S_0, \dots, S_N with $S_0 = 1$ such that (6.19) holds for all $x_1, y_1, \dots, x_N, y_N$. Because no weakly N -calibrated probabilistic predictor exists, no N -calibrated or strongly N -calibrated probabilistic predictor exists. But as we will see in the next section, there are multiprobability predictors that satisfy an analogue of strong N -calibration.

6.2 On-line multiprobability prediction

We now generalize the protocol for probabilistic prediction: instead of being required to give a single probability distribution for the forthcoming label, Predictor can give several. Multiprobability prediction in this sense will be our topic for the remainder of the chapter.

We call a measurable strategy for Predictor in our multiprobability prediction protocol a *multiprobability predictor*. (Essentially, the notion of a multiprobability predictor is a simplified version of that of a probability estimator

in §5.2.) In this section, we spell out the protocol and state our criterion for validity for a multiprobability predictor. In the next section, we describe the class of multiprobability predictors that we call Venn predictors and state a theorem showing that they satisfy this criterion.

The on-line protocol

Our generalization of the probabilistic prediction protocol is very simple. Instead of asking Predictor to give a single probability distribution $p_n \in \mathbf{P}(\mathbf{Y})$, we ask him to give a set of probability distributions $P_n \subseteq \mathbf{P}(\mathbf{Y})$. So play goes like this:

```
MULTIPROBABILITY PREDICTION
FOR  $n = 1, 2, \dots, N$ :
    Reality announces  $x_n \in \mathbf{X}$ ;
    Predictor announces  $P_n \subseteq \mathbf{P}(\mathbf{Y})$ ;
    Reality announces  $y_n \in \mathbf{Y}$ 
END FOR.
```

In order for the predictions P_n to be useful, their constituent probability distributions should not be too different from one another. We would hope that for large n and each $y \in \mathbf{Y}$, the probabilities $p\{y\}$ given by different $p \in P_n$ will all fall within a small interval, so that we can say that Predictor has given an approximate probability for y .

Remark The measurability of a multiprobability predictor means that the set

$$\{(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, p) : p \in F(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)\}$$

is a measurable subset of $\mathbf{Z}^{n-1} \times \mathbf{X} \times \mathbf{P}(\mathbf{Y})$ for all $n = 1, \dots, N$ (cf. (5.3) on p. 133). However, for the strategies that interest us, the Venn predictors we describe in the next section, P_n always consists of a finite number of probability distributions on the finite set \mathbf{Y} , the number being at most the number of elements in \mathbf{Y} . Therefore, the reader might wish to replace the assumption that P_n is a set of probability distributions on \mathbf{Y} with the assumption that it is a list of K probability distributions on \mathbf{Y} , repetitions being permitted. This means replacing “ $P_n \subseteq \mathbf{P}(\mathbf{Y})$ ” by “ $P_n \in (\mathbf{P}(\mathbf{Y}))^K$ ” in the protocol.

We use the same notions of *game martingale* and *game supermartingale* as in probabilistic prediction (see p. 150), except that we extend the domain of definition for a game martingale or supermartingale G from sequences of the form $p_1, y_1, \dots, p_n, y_n$, where p_1, \dots, p_n are single probability distributions on \mathbf{Y} , to sequences of the form $P_1, y_1, \dots, P_n, y_n$, where P_1, \dots, P_n are sets of probability distributions on \mathbf{Y} , by

$$G(P_1, y_1, \dots, P_n, y_n) := \inf_{p_1 \in P_1, \dots, p_n \in P_n} G(p_1, y_1, \dots, p_n, y_n) . \quad (6.20)$$

This definition is motivated by our intuitive picture of multiprobability forecasting, in which Predictor only claims that some p in P_n describes y_n well, and so we can claim to reject the multiprobability forecasts P_1, \dots, P_n if and only if we have rejected all probability forecasts $(p_1, \dots, p_n) \in P_1 \times \dots \times P_n$.

We will use the notions of a calibration/reversible martingale/supermartingale introduced in the previous section. If a calibration martingale or supermartingale G starts with $G(\square) = 1$ and ends with $G(P_1, y_1, \dots, P_N, y_N)$ very large, then we may reject the calibration of Predictor's P_1, \dots, P_N .

Validity

We call a multiprobability predictor F *N-valid* if for any exchangeable probability distribution P on \mathbf{Z}^N and any nonnegative reversible game supermartingale G with $G(\square) = 1$, there exists a P -supermartingale S_0, \dots, S_N with $S_0 = 1$ such that

$$G(F(x_1), y_1, \dots, F(x_1, y_1, \dots, x_N), y_N) \leq S_N(x_1, y_1, \dots, x_N, y_N)$$

for all $x_1, y_1, \dots, x_N, y_N$. This is analogous to the condition of being strongly N -calibrated for probabilistic predictors (p. 156).

6.3 Venn predictors

In this section we formally define Venn predictors, which were described informally in the introduction, and we show that they are valid multiprobability predictors.

We begin with the concept of a *taxonomy* (or, more fully, *Venn taxonomy*). This is a sequence A_n , $n = 1, \dots, N$, where each A_n is a measurable finite partition of the space $\mathbf{Z}^{(n-1)} \times \mathbf{Z}$. (A finite partition is said to be *measurable* if each element of the partition is measurable; for simplicity, we only consider finite partitions here.) As usual, we write $A_n(\omega)$ for the element of the partition A_n that contains $\omega \in \mathbf{Z}^{(n-1)} \times \mathbf{Z}$. For every taxonomy A_1, A_2, \dots, A_N , we will define a Venn predictor.

Having chosen a taxonomy A_1, A_2, \dots, A_N , consider a label $y \in \mathbf{Y}$, and consider the situation in the multiprobability prediction protocol where Reality has made the moves $x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n$ and Predictor is about to make his move P_n . Write (as usual) z_i for (x_i, y_i) and (just for the moment) z_n for (x_n, y) . Then partition the bag $\{z_1, \dots, z_n\}$ into categories, assigning z_i and z_j to the same category if and only if

$$A_n(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}, z_i) = A_n(\{z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n\}, z_j) .$$

The category T containing $z_n = (x_n, y)$ is nonempty, because it contains at least this one element. Let p_y be the empirical probability distribution of the labels in this category T :

$$p_y\{y'\} := \frac{|\{(x^*, y^*) \in T : y^* = y'\}|}{|T|},$$

this is a probability distribution on \mathbf{Y} . The *Venn predictor* determined by the taxonomy is the multiprobability predictor $P_n := \{p_y : y \in \mathbf{Y}\}$. The set P_n consists of between one and K distinct probability distributions on \mathbf{Y} , where $K = |\mathbf{Y}|$.

There are many Venn predictors, one for each taxonomy. Some will be more efficient than others on particular data sets. But all share one virtue:

Theorem 6.6. *Every Venn predictor is an N -valid multiprobability predictor².*

In order to emphasize the applicability of our theory to finite data sets, we have fixed the horizon N . But the moves recommended by a Venn predictor do not depend on the particular horizon N . The moves recommended on trial n depend only on the partition A_n , and they do not change if we increase N by extending the sequence of partitions A_1, A_2, \dots, A_N .

The problem of the reference class

In choosing the partitions that determine a Venn predictor, we face the dilemma that is often called the “problem of the reference class”. We want the categories into which we divide the examples to be large, in order to have a reasonable sample size for estimating the probabilities. But we also want them to be small and homogeneous. According to Kılıç (2001), John Venn was the first to formulate and analyze this problem with due philosophical depth. To see how the problem of the reference class comes into our picture, it suffices to consider the case where \mathbf{Y} is binary: $\mathbf{Y} = \{0, 1\}$.

First consider the *Bernoulli problem*: we observe only the successive labels y_1, \dots, y_N , not preceded by objects; as usual in the binary case, p_n will be the probability Predictor announces for the event $y_n = 1$. In this problem, we want to make a probabilistic prediction for each new label y_n in light of the previous labels y_1, \dots, y_{n-1} . The most naive probabilistic prediction is $p_n = k/(n-1)$, where k is the number of 1s among the first $n-1$ labels and it is assumed that $n > 1$. There are other possibilities, such as *Laplace's rule of succession*, $p_n = (k+1)/(n+1)$. But when n is large, $k/(n-1)$ and $(k+1)/(n+1)$ are close to each other and to any other reasonable choice for p_n . The only natural Venn predictor agrees with this consensus: since there are no objects x_n , we take each A_n to be the partition that puts all the examples in the

²This being a special case of Theorem 9.1 (p. 224) in Chap. 9, we do not provide a proof in this chapter.

same category, and this produces the Venn predictor $P_n = \{k/n, (k+1)/n\}$; in particular, the convex hull $\text{co } P_n$ contains both $k/(n-1)$ and $(k+1)/(n+1)$. Notice that since there are only n examples, the diameter $1/n$ of P_n for this Venn predictor is the smallest achievable.

Now suppose $|\mathbf{X}| > 1$ and the x_n vary a great deal, so that the examples z_1, \dots, z_{n-1} are quite heterogeneous. In this case the probabilistic prediction $k/(n-1)$ for y_n seems too crude when n is large. A more reasonable predictor would take into account only objects x_i that are similar, in a suitable sense, to x_n . The classical approach, described on p. 143, is:

- Split the available objects x_1, \dots, x_{n-1} into a number of categories.
- Output k'/n' as the probability that $y_n = 1$, where n' is the number of objects among x_1, \dots, x_{n-1} in the same category as x_n and k' is the number of objects among those n' that are labeled as 1.

This is where we face the problem of the reference class.

The procedure that produces a Venn predictor is a simple modification of the classical procedure:

- Consider the two possible completions of the known data

$$(z_1, \dots, z_{n-1}, x_n) = ((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), x_n) :$$

in one (the 0-completion) x_n is assigned label 0, and in the other (the 1-completion) x_n is assigned label 1.

- In each completion, split all examples $z_1, \dots, z_{n-1}, (x_n, y)$ into a number of categories, so that the split does not depend on the order of examples ($y = 0$ for the 0-completion and $y = 1$ for the 1-completion).
- In each completion, output k'/n' as the probability that $y_n = 1$, where n' is the number of examples among $z_1, \dots, z_{n-1}, (x_n, y)$ in the same category as (x_n, y) and k' is the number of examples among those n' that are labeled as 1.

The new procedure differs from the classical one in two salient ways: (1) we can use the old labels as well as the old objects in dividing the old examples into categories, and (2) we now have two predicted probabilities instead of one. The first difference offers an advantage with no obvious disadvantage: we have greater flexibility in how we divide the old examples into categories. The second difference, having two probabilities for $y_n = 1$ rather than one, might be considered a disadvantage for the new procedure, but it can also be considered an advantage. If the two probabilities are quite different, then the uncertainty in the probability can be considered substantial, and the new procedure makes this uncertainty visible rather than hiding it.

Of course, the most important advantage of the new procedure is its validity: the procedure is automatically calibrated in a quite satisfactory sense, no matter how we choose the “reference classes”. The reference class problem that remains can be considered an issue for efficiency. We must balance

two kinds of inefficiency. Too many categories in our partition is a kind of overfitting, and it is punished by a large diameter for the multiprobability prediction. Too few categories is a kind of underfitting, and it is punished by predictions that are not close enough to zero or one.

Empirical results

In this section, we report how a natural Venn predictor performs on the USPS data set (see §B.1).

The taxonomy that defines the Venn predictor is based on the 1-nearest neighbor algorithm. Since the data set is relatively small (9298 examples in total), we make the taxonomy very coarse: two examples are assigned to the same category if their nearest neighbors have the same label. This produces 10 categories. The distance between two examples is defined as the Euclidean distance between their objects (16×16 matrices of pixels, represented as points in \mathbb{R}^{256}).

The algorithm processes the n th object x_n as follows. First it creates the 10×10 matrix A whose entry $A_{i,j}$, $i, j = 0, \dots, 9$, is computed by assigning i to x_n as label and finding the fraction of examples labeled j among the examples in the bag $\{z_1, \dots, z_{n-1}, (x_n, i)\}$ belonging to the same category as (x_n, i) . The *quality* of a column of this matrix is its minimum entry. Choose a column (called the *best* column) with the highest quality; let the best column be j_{best} . Output j_{best} as the prediction and output

$$\left[\min_{i=0,\dots,9} A_{i,j_{\text{best}}}, \max_{i=0,\dots,9} A_{i,j_{\text{best}}} \right]$$

as the interval for the probability that this prediction is correct. If the latter interval is $[a, b]$, the complementary interval $[1 - b, 1 - a]$ is called the *error probability interval*. We say that this procedure makes an error when predicting y_n if $y_n \neq j_{\text{best}}$.

In Fig. 6.1 we show the following three curves: the cumulative error curve

$$E_n := \sum_{i=1}^n \text{err}_i ,$$

where $\text{err}_i = 1$ if an error is made at trial i and $\text{err}_i = 0$ otherwise; the *cumulative lower error probability curve*

$$L_n := \sum_{i=1}^n l_i$$

and the *cumulative upper error probability curve*

$$U_n := \sum_{i=1}^n u_i ,$$

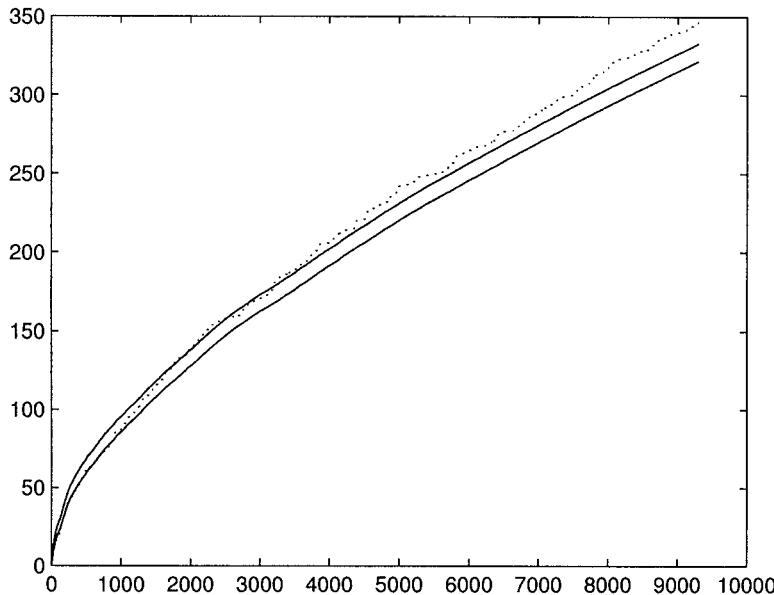


Fig. 6.1. On-line performance of the 1-nearest neighbor Venn predictor on the USPS data set (9298 hand-written digits, randomly permuted). The dotted line shows the cumulative number of errors E_n and the solid ones the cumulative upper and lower error probability curves U_n and L_n . In this particular experiment, the mean error E_N/N is 4.25% and the mean probability interval $(1/N)[L_N, U_N]$ is [4.07%, 4.19%], where $N = 9298$ is the size of the data set

where $[l_i, u_i]$ is the error probability interval output by the algorithm at trial i for the label y_i ; the values E_n , L_n and U_n are plotted against n . The plot confirms that the error probability intervals are well calibrated.

Probabilities vs. p-values

In the earlier chapters we saw that it is possible to produce valid and asymptotically optimal p-values and to put them to practical use in confidence prediction. They do have disadvantages, however, relative to probabilities: their interpretation is less direct than that of probabilities, and many people confuse them with probabilities. These disadvantages weigh so heavily with some authors that they counsel against any use of p-values (see, e.g., Berger and Delampady 1987).

The multiprobabilities defined in this chapter seem to address some of these concerns. We have already mentioned that a family of prediction sets Γ^ϵ output by a confidence predictor can be usefully summarized by reporting the confidence (3.66) (p. 96) in this prediction; it is easy to see that the confidence can also be expressed as one minus the second largest among the

p-values computed for all the potential labels for the new object. In some ways confidence is analogous to the probability that the simple prediction corresponding to the maximal p-value is correct; the difference, however, is very important.

Suppose, for example, that a Venn predictor outputs $p_n \approx 99\%$ in the binary case. Then we expect $y_n = 1$ and we know that we will be wrong in about 1% of similar examples, those where our prediction is close to 99%. This is different from what we know when a valid confidence predictor outputs $y_n = 1$ with confidence 99%. With a valid confidence predictor, we can only assert that the frequency with which we will be wrong for similar or more extreme examples (examples where our confidence is at least 99%) will be close to or less than 1%. It is the inclusion of the clause “or more extreme” that some authors find unconvincing. Another important difference is that in the case of confidence predictor the frequency is taken over the full data sequence, not just over examples predicted confidently.

6.4 A universal Venn predictor

The following result asserts the existence of a universal Venn predictor. As the proof (given in the next section) shows, such a predictor can be constructed quite easily, using the histogram approach to probability estimation (Devroye et al. 1996).

Theorem 6.7. *Suppose \mathbf{X} is a Borel space and \mathbf{Y} is finite. Let Q be a probability distribution on \mathbf{Z} with regular conditional probabilities $Q_{\mathbf{Y}|\mathbf{X}}(\cdot | \cdot)$. There exists a Venn predictor such that, if the examples are generated from Q^∞ ,*

$$\sup_{p \in P_n} \rho(p, Q_{\mathbf{Y}|\mathbf{X}}(\cdot | x_n)) \rightarrow 0 \quad (n \rightarrow \infty) \quad (6.21)$$

in probability, where P_n are the multiprobabilities produced by the Venn predictor and ρ is the variation distance,

$$\rho(p, q) := \sum_{y \in \mathbf{Y}} |p\{y\} - q\{y\}| .$$

This theorem can be interpreted by saying that some Venn predictors have asymptotically optimal efficiency. We proved a similar result for p-values in Chap. 3. Our current result is much easier to prove but is weaker in form: it only asserts convergence in probability. Whether there exists a Venn predictor for which (6.21) holds almost surely is an open question.

6.5 Proofs

Proof of Theorem 6.5

Assume, without loss of generality, that the object are absent ($|\mathbf{X}| = 1$) and $\mathbf{Y} = \{0, 1\}$ (remember that $|\mathbf{Y}| \geq 2$ in this book). We will prove that for

any probabilistic predictor F there exist a probability distribution Q on \mathbf{Y} and a calibration event E such that the Q^N -probability that the sequence $p_1y_1 \dots p_Ny_N$ of Predictor and Reality's moves belongs to E exceeds $\bar{\mathbb{P}}(E)$; $\bar{\mathbb{P}}(E)$ will be understood in the sense of $\bar{\mathbb{P}}^{\text{game}}(E)$, since $\bar{\mathbb{P}}^{\text{meas}}(E) \leq \bar{\mathbb{P}}^{\text{game}}(E)$.

Fix a probabilistic predictor F . Define sequences $y_1^* \dots y_N^* \in \{0, 1\}^N$ and $p_1^* \dots p_N^* \in [0, 1]^N$ inductively as follows:

$$y_n^* := \begin{cases} 1 & \text{if } p_n^* := F(y_1^*, \dots, y_{n-1}^*) \leq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

(cf. Dawid 1985). Let E be the game event consisting of $p_1^*y_1^* \dots p_N^*y_N^*$ and all its permutations. (As usual in the binary case, we identify a probability distribution p on $\{0, 1\}$ with the number $p\{1\} \in [0, 1]$.) There are two possible cases:

Not all p_n^* equal 0.5: define Q as the Bernoulli distribution on $\{0, 1\}$ with parameter 0.5. The Q^N -probability that the sequence $p_1y_1 \dots p_Ny_N$ of Predictor and Reality's moves belongs to E will be at least 2^{-N} (this is the probability that Reality generates exactly $y_1^* \dots y_N^*$) and $\bar{\mathbb{P}}^{\text{game}}(E)$ will be less than 2^{-N} .

All p_n^* equal 0.5: define Q as the Bernoulli distribution on $\{0, 1\}$ with parameter 1: $Q\{1\} = 1$. The Q^N -probability that the sequence $p_1y_1 \dots p_Ny_N$ of Predictor and Reality's moves belongs to E will now be equal to 1 (since all y_n^* are now 1) and $\bar{\mathbb{P}}^{\text{game}}(E)$ will be precisely 2^{-N} .

In both cases the Q^N -probability that the sequence $p_1y_1 \dots p_Ny_N$ of Predictor and Reality's moves belongs to E exceeds $\bar{\mathbb{P}}^{\text{game}}(E)$.

Equivalence of the two definitions of upper probability

The equivalence between (6.5) and (6.10) (pp. 149 and 151) is analogous to Ville's theorem (see §A.6), and a similar result is proved in Shafer 1996a. Unfortunately, to avoid technical difficulties we have to impose two assumptions of finiteness, but these assumptions are not restrictive in the finite world of applications (cf. p. 133).

Fix a finite set $D \subseteq \mathbf{P}(\mathbf{Y})$ and modify (6.4) (p. 148) to

$$\Pi_D := (D \times \mathbf{Y})^N.$$

Lemma 6.8. Suppose the object space \mathbf{X} is finite. For all $E \subseteq \Pi_D$,

$$\bar{\mathbb{P}}^{\text{meas}}(E) = \bar{\mathbb{P}}^{\text{game}}(E).$$

Proof. We are required to prove that

$$\begin{aligned} \sup_R R \left\{ x_1y_1 \dots x_Ny_N : p_1^R y_1 \dots p_N^R y_N \in E \right\} \\ = \inf \left\{ G(\square) : G(p_1, y_1, \dots, p_N, y_N) \geq 1, \right. \\ \left. \forall (p_1, y_1, \dots, p_N, y_N) \in E \right\}, \quad (6.22) \end{aligned}$$

R ranging over the probability distributions on \mathbf{Z}^N and G ranging over the nonnegative game supermartingales; p_i^R is the regular conditional distribution under R for the different possible values of y_i given z_1, \dots, z_{i-1} and x_i .

The part “ \leq ” of (6.22) follows from the fact that

$$R \{ G(p_1^R, y_1, \dots, p_N^R, y_N) \geq 1 \} \leq G(\square) \quad (6.23)$$

for all nonnegative game supermartingales G . Since

$$S_n := G(p_1^R, y_1, \dots, p_n^R, y_n)$$

is an R -supermartingale (by Lemma A.2 on p. 281), (6.23) is a special case of Doob’s inequality (p. 285). Therefore, this part holds in general and does not require the assumptions that \mathbf{X} is finite and that p_n are constrained to D .

Equality (6.22) follows from the fact that both its sides equal

$$\begin{aligned} & \sup_{p_1 \in D} \int_{\mathbf{Y}} \dots \sup_{p_{N-1} \in D} \int_{\mathbf{Y}} \sup_{p_N \in D} \int_{\mathbf{Y}} \\ & \quad \mathbb{I}_E(p_1, y_1, \dots, p_{N-1}, y_{N-1}, p_N, y_N) \\ & \quad p_N(dy_N) p_{N-1}(dy_{N-1}) \dots p_1(dy_1). \end{aligned}$$

This can be seen by setting

$$\begin{aligned} f(p_1, y_1, \dots, p_n, y_n) := & \sup_{p_{n+1} \in D} \int_{\mathbf{Y}} \dots \sup_{p_N \in D} \int_{\mathbf{Y}} \\ & \quad \mathbb{I}_E(p_1, y_1, \dots, p_N, y_N) p_N(dy_N) \dots p_{n+1}(dy_{n+1}) \end{aligned}$$

and showing that, for all n and all $p_1, y_1, \dots, p_n, y_n$,

$$\begin{aligned} f(p_1, y_1, \dots, p_n, y_n) = & \\ & \sup_R R \left\{ x_{n+1}^* y_{n+1}^* \dots x_N^* y_N^* : p_{n+1}^R y_{n+1}^* \dots p_N^R y_N^* \in E \right\} \end{aligned}$$

(R ranges over the probability distributions on \mathbf{Z}^{N-n} and p_i^R , for $i = n + 1, \dots, N$, is the conditional probability distribution under R for the different possible values of the i th label given that the first $i - 1$ examples and the i th object are $x_1, y_1, \dots, x_n, y_n, x_{n+1}^*, y_{n+1}^*, \dots, x_{i-1}^*, y_{i-1}^*, x_i^*$) and

$$\begin{aligned} f(p_1, y_1, \dots, p_n, y_n) = & \\ & \inf \{ G(p_1, y_1, \dots, p_n, y_n) : G(p_1, y_1, \dots, p_N, y_N) \geq 1, \\ & \quad \forall (p_{n+1}, y_{n+1}, \dots, p_N, y_N) \in E \} \end{aligned}$$

(with G ranging over the nonnegative game supermartingales) by induction on $n = N, N - 1, \dots, 0$. \square

Proof of Theorem 6.7

Since \mathbf{X} is a Borel space, we assume, without loss of generality, that $\mathbf{X} = [0, 1]$. For each $n = 1, 2, \dots$ consider the partition of the interval $[0, 1]$ into bins B_k , which will now be denoted $B_{n,k}$, given by (6.15) (p. 153); the number of bins $K = K_n$ is now allowed to depend on n . We will be interested in the case where $K_n \rightarrow \infty$ but $K_n/n \rightarrow 0$ as $n \rightarrow \infty$.

Let (x_i, y_i) , $i = 1, 2, \dots$, be the examples output by Reality. Define, for every $(x, y) \in \mathbf{Z}$,

$$Q_n(y | x) := \frac{N_n(x, y)}{N_n(x)}, \quad Q_n^*(y | x) := \frac{N_n(x, y)}{nQ_{\mathbf{X}}(B_n(x))},$$

where $B_n(x)$ is the bin in the n th partition (consisting of the bins $B_{n,k}$, $k = 1, \dots, K_n$) containing x , $N_n(x)$ is the number of $i = 1, \dots, n$ such that $x_i \in B_n(x)$, $N_n(x, y)$ is the number of $i = 1, \dots, n$ such that $x_i \in B_n(x)$ and $y_i = y$, and the uncertainty 0/0 is resolved to, say, $1/|\mathbf{Y}|$.

We will need the following analog of Lemma 3.9 (p. 83).

Lemma 6.9. *Suppose $K_n \rightarrow \infty$, $K_n = o(n)$, and $\mathbf{Y} = \{0, 1\}$. For any $\delta > 0$ and large enough n ,*

$$\mathbb{P} \left\{ \int |Q(1 | x) - Q_n^*(1 | x)| Q_{\mathbf{X}}(dx) > \delta \right\} \leq e^{-n\delta^2/8}$$

where the outermost probability distribution $\mathbb{P} = Q^\infty$ generates the examples (x_i, y_i) , which determine the empirical distributions Q_n and “semi-empirical distributions” Q_n^* .

Proof. See Devroye et al. 1996, Theorem 9.4 and the second displayed equation on p. 139. \square

As in Chap. 3 (see the proof of Lemma 3.10), we obtain the following lemma for the multilabel case.

Lemma 6.10. *Suppose $K_n \rightarrow \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n ,*

$$\mathbb{P} \left\{ Q_{\mathbf{X}} \left\{ x : \max_{y \in \mathbf{Y}} |Q_n^*(y | x) - Q(y | x)| > \delta \right\} > \delta \right\} \leq e^{-\delta^* n}. \quad (6.24)$$

This lemma implies the analogous statement (Corollary 6.12) for the empirical distributions Q_n , but we need an intermediate step.

Lemma 6.11. *Suppose $K_n \rightarrow \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n ,*

$$\mathbb{P} \left\{ Q_{\mathbf{X}} \left\{ x : \left| \frac{N_n(x)/n}{Q_{\mathbf{X}}(B_n(x))} - 1 \right| > \delta \right\} > \delta \right\} \leq e^{-\delta^* n}.$$

Proof. Replacing $\max_{y \in \mathbf{Y}}$ by $\sum_{y \in \mathbf{Y}}$ in (6.24), we obtain

$$\mathbb{P} \left\{ Q_{\mathbf{X}} \left\{ x : \sum_{y \in \mathbf{Y}} |Q_n^*(y | x) - Q(y | x)| > |\mathbf{Y}| \delta \right\} > \delta \right\} \leq e^{-\delta^* n} ;$$

it remains to notice that

$$\begin{aligned} \sum_{y \in \mathbf{Y}} |Q_n^*(y | x) - Q(y | x)| &\geq \left| \sum_{y \in \mathbf{Y}} Q_n^*(y | x) - \sum_{y \in \mathbf{Y}} Q(y | x) \right| \\ &= \left| \frac{N_n(x)/n}{Q_{\mathbf{X}}(B_n(x))} - 1 \right|. \quad \square \end{aligned}$$

The two preceding lemmas immediately give

Corollary 6.12. *Suppose $K_n \rightarrow \infty$ and $K_n = o(n)$. For any $\delta > 0$ there exists a $\delta^* > 0$ such that, for large enough n ,*

$$\mathbb{P} \left\{ Q_{\mathbf{X}} \left\{ x : \max_{y \in \mathbf{Y}} |Q_n(y | x) - Q(y | x)| > \delta \right\} > \delta \right\} \leq e^{-\delta^* n} .$$

The following result is proved in Devroye et al. 1996 (Theorem 6.2 and its proof):

Lemma 6.13. *Suppose $K_n \rightarrow \infty$ and $K_n = o(n)$. For any constant C ,*

$$Q_{\mathbf{X}} \{x : N_n(x) > C\} \rightarrow 1$$

in probability as $n \rightarrow \infty$.

Now it is easy to prove Theorem 6.7. Consider the Venn predictor determined by the taxonomy in which $A_n(D, (x, y))$ consists of all $(D', (x', y'))$ such that x and x' are in the same bin $B_{n,k}$ (so that $A_n(D, (x, y))$ does not depend on the bag $D \in \mathbf{Z}^{(n-1)}$ or the label $y \in \mathbf{Y}$). It suffices to show that

$$\text{diam}(P_n) \rightarrow 0 \tag{6.25}$$

and

$$\rho(Q_n(\cdot | x_n), Q(\cdot | x_n)) \rightarrow 0 \tag{6.26}$$

in probability as $n \rightarrow \infty$ (remember that, by the definition of Venn predictor, $Q_n(\cdot | x_n) \in P_n$). But this is simple: (6.25) follows from Lemma 6.13 and (6.26) follows from Corollary 6.12.

6.6 Bibliographical remarks

Testing

The theory of testing statistical hypotheses based on Cournot's principle has a very long history, going back at least to Arbuthnott 1710–1712, and the literature devoted to this topic is vast. Testing using martingales is also popular: for example, martingales (in the form of probability ratios) are widely used in sequential analysis for this purpose. However, even when testing is done using martingales, the basic principle is usually still Cournot's (see, e.g., Wald 1947, Wald and Wolfowitz 1948); the value taken by a nonnegative martingale starting from one is rarely interpreted as measuring the weight of evidence found against the statistical hypothesis. The martingale approach to testing free of Cournot's principle is discussed in Vovk 1993, Shafer and Vovk 2001.

Frequentist probability

In this chapter we have been concerned with probabilities that are valid in the sense of agreeing with the observed frequencies (cf. (6.2) on p. 147). John Venn was one of the first writers on frequentist probability; see his book Venn 1866. A major figure in the area was Richard von Mises, who connected the frequentist notion of probability with the principle of the excluded gambling system (see Mises 1919, Mises 1928); the latter lead to Ville's notion of martingale (Ville 1939).

Beyond exchangeability

In Chaps. 2–6 we assumed that all examples output by Reality are exchangeable. This is a strong assumption, but it is standard in machine learning (where the even stronger assumption of randomness is usually made). In this chapter we discuss how to test and how to relax this assumption.

Section 7.1 considers the problem of testing the exchangeability assumption. This problem is important in its own right and also needed in §7.2, where we discuss modeling a “stochastically dynamic” environment.

Intuitively, the exchangeability assumption means that Reality is stochastically static. This is especially clear if we recast it as the randomness assumption (assuming that the example space \mathbf{Z} is Borel and appealing to de Finetti’s theorem of §A.5); the examples are then generated from the same stochastic mechanism independently. In §7.2 we assume, instead, that there is a static “random core” on which a less complicated dynamic structure is superimposed. In §7.3 we briefly consider another relaxation of the assumption of randomness: it is only assumed that certain subsequences of the full data sequence output by Reality satisfy this assumption. It is possible to combine relaxations considered in §§7.2 and 7.3, but this is straightforward and we do not discuss the details.

7.1 Testing exchangeability

This section discusses on-line ways of monitoring the strength of evidence against the assumption of exchangeability. Such on-line monitoring is often a wise thing to do even if the exchangeability assumption is tentatively accepted. We already saw in the previous chapter that a convenient and natural on-line way of testing statistical hypotheses is provided by nonnegative supermartingales. Since, however, this chapter is conceptually simpler than the previous one, we made these two chapters independent of each other.

We will start this section by introducing the notion of exchangeability supermartingales, which are in effect on-line procedures for detecting deviations

from exchangeability. Intuitively, nonnegative exchangeability supermartingales are betting schemes that never risk bankruptcy and do not benefit the gambler under the hypothesis of exchangeability. We then construct some specific exchangeability supermartingales from conformal transducers, introduced in §2.5. Finally, we will report experimental results showing the performance of those exchangeability supermartingales on the USPS benchmark data set of hand-written digits (known to be somewhat heterogeneous; see Appendix B); one of them multiplies the initial capital by more than 10^{18} ; this could be expressed in statistical terms by saying that the hypothesis of exchangeability is rejected at the significance level 10^{-18} .

Exchangeability supermartingales

In this section we set up our basic framework, defining the fundamental notion of exchangeability supermartingale and the closely related notion of randomized exchangeability martingale. In our standard learning protocol, Reality outputs examples z_1, z_2, \dots , each of which consists of two parts, an object and its label. In the theoretical considerations of this chapter, however, we will not use this additional structure; therefore, the example space \mathbf{Z} is not assumed to be a Cartesian product $\mathbf{X} \times \mathbf{Y}$.

We are interested in testing the hypothesis of exchangeability *on-line*: after observing each new example z_n we would like to have a number M_n reflecting the strength of evidence found against the hypothesis. Let us first consider testing the simple hypothesis that z_1, z_2, \dots are generated from a probability distribution P on \mathbf{Z}^∞ . We say that a sequence of random variables M_0, M_1, \dots is a *P-supermartingale* if, for all $n = 0, 1, \dots$, M_n is a measurable function of z_1, \dots, z_n (in particular, M_0 is a constant) and

$$M_n \geq \mathbb{E}(M_{n+1} | M_1, \dots, M_n) \quad \text{a.s.}, \quad (7.1)$$

where \mathbb{E} refers to the expected value in the probability space in which z_1, z_2, \dots are generated from P . If $M_0 = 1$ and $\inf_n M_n \geq 0$, M_n can be regarded as the capital process of a player who starts from 1, never risks bankruptcy, at the beginning of each trial n places a fair (cf. (7.1)) bet on the z_n to be chosen by Reality, and maybe sometimes throws money away (since (7.1) is an inequality). If such a supermartingale M ever takes a large value, our belief in P is undermined; this intuition is formalized by Doob's inequality (see §A.6), which implies

$$P\{(z_1, z_2, \dots) : \exists n : M_n \geq C\} \leq 1/C, \quad (7.2)$$

where C is an arbitrary positive constant.

When testing a *composite hypothesis* \mathcal{P} (i.e., a family of probability distributions on \mathbf{Z}^∞), we will use *\mathcal{P} -supermartingales*, i.e., sequences of random variables M_0, M_1, \dots which are *P*-supermartingales for all $P \in \mathcal{P}$ simultaneously. We are primarily interested in the family \mathcal{P} consisting of all exchangeable probability distributions P on \mathbf{Z}^∞ ; in this case we will say *exchangeability supermartingales* to mean *\mathcal{P} -supermartingales*.

Remark If \mathcal{P} is the set of all power probability distributions Q^∞ , Q ranging over the probability distributions on \mathbf{Z} , \mathcal{P} -supermartingales are called *randomness supermartingales*. De Finetti's theorem (see §A.5) and the fact that Borel spaces are closed under countable products (see, e.g., Schervish 1995, Lemma B.41) imply that each exchangeable distribution P on \mathbf{Z}^∞ is a mixture of power distributions Q^∞ provided \mathbf{Z} is Borel. By Property 3 (p. 280) of conditional probability distributions the notions of randomness and exchangeability supermartingales coincide in the Borel case. But even without the assumption that \mathbf{Z} is Borel, all exchangeability supermartingales are randomness supermartingales.

Another useful notion is that of *randomized exchangeability martingales*; these are sequences of measurable functions $M_n(z_1, \tau_1, \dots, z_n, \tau_n)$ (each example z_n is extended by adding a random number $\tau_n \in [0, 1]$) such that, for any exchangeable probability distribution P on \mathbf{Z}^∞ ,

$$M_n = \mathbb{E}(M_{n+1} | M_1, \dots, M_n) \quad \text{a.s. ,} \quad (7.3)$$

\mathbb{E} referring to the expected value in the probability space in which z_1, z_2, \dots and τ_1, τ_2, \dots are generated from P and \mathbf{U}^∞ (remember that \mathbf{U} is the uniform distribution on $[0, 1]$) independently.

Remark An exchangeability martingale is defined as an exchangeability supermartingale such that (7.1) holds as equality for any exchangeable P , and the notion of randomized exchangeability supermartingale is obtained by relaxing the “=” in (7.3) to “ \geq ”. We do not need these notions, however: the notion of exchangeability martingale is too restrictive and that of randomized exchangeability supermartingale is unnecessarily wide (our goals can be achieved already with randomized exchangeability martingales).

Remark In our definitions of martingale (7.3) and supermartingale (7.1) we follow Doob 1953, §II.7 and the beginning of §VII.1. (Doob, however, did not use the term “supermartingale”; for details, see Snell 1997, p. 307.) A more modern approach (cf. Shiryaev 1996, Shafer and Vovk 2001; introduced already in Doob 1953) would be to replace the condition “ $| M_1, \dots, M_n$ ” in (7.3) and (7.1) by “ $| \mathcal{F}_n$ ”, where \mathcal{F}_n is the σ -algebra generated by z_1, \dots, z_n in the case of (7.1) and $z_1, \tau_1, \dots, z_n, \tau_n$ in the case of (7.3) (i.e., \mathcal{F}_n represents all information available by the end of trial n). To see how restrictive conditions (7.3) and (7.1) are, notice that the notions of randomized exchangeability martingale and exchangeability supermartingale become trivial when this apparently small change is made: if the example space \mathbf{Z} is Borel, the latter will be decreasing processes ($M_0 \geq M_1 \geq \dots$) and the former will only bet on the random numbers τ_1, τ_2, \dots .

Power supermartingales and the simple mixture

We know from §2.5 that the p-values p_1, p_2, \dots output by a smoothed conformal transducer are independent and distributed uniformly in $[0, 1]$. They can

be used for constructing exchangeability supermartingales and randomized exchangeability martingales.

Since $\int_0^1 \epsilon p_i^{\epsilon-1} dp = 1$ for $\epsilon > 0$, the random variables

$$M_n^{(\epsilon)} := \prod_{i=1}^n (\epsilon p_i^{\epsilon-1}) = \epsilon^n \left(\prod_{i=1}^n p_i \right)^{\epsilon-1}, \quad (7.4)$$

where p_i are the p-values output by a smoothed conformal transducer, will form a nonnegative randomized exchangeability martingale with initial value 1; this family of martingales, indexed by $\epsilon \in [0, 1]$, will be called the *power martingales* (notice that $M_n^{(0)}$ is different from the other power martingales in that $M_0^{(0)} \neq 1$). To eliminate the dependence on ϵ , we may use the randomized exchangeability martingale

$$M_n := \int_0^1 M_n^{(\epsilon)} d\epsilon, \quad (7.5)$$

which is called the *simple mixture* of $M_n^{(\epsilon)}$.

If p_1, p_2, \dots are produced by a deterministic conformal transducer, the random variables (7.4) will form a nonnegative exchangeability supermartingale with initial value 1 (unless $\epsilon = 0$), and this family of supermartingales is called the *power supermartingales*. The simple mixture (7.5) will then be an exchangeability supermartingale.

Remark There is a simple way to compute the values M_n of the simple mixture (7.5). Denoting $b := -\ln(p_1 \cdots p_n)$, where p_i are the p-values output by a (smoothed) conformal transducer, we find:

$$M_n = \int_0^1 \epsilon^n e^{-b(\epsilon-1)} d\epsilon = e^b b^{-n-1} \int_0^b t^n e^{-t} dt = e^b b^{-n-1} \gamma(n+1, b),$$

where

$$\gamma(a, b) := \int_0^b t^{a-1} e^{-t} dt, \quad a > 0, \quad b \geq 0,$$

is one of the definitions of the incomplete gamma function.

All experiments described in this section are performed on the full USPS data set with no pre-processing of images, as described in Appendix B. Our goal will be to detect deviations from exchangeability for this data set.

We saw in Chap. 3 that a nearest neighbors smoothed conformal predictor provides a universally optimal, in an asymptotic sense, on-line algorithm for prediction under the assumption of exchangeability. On the empirical side, we also saw that a 1-nearest neighbor conformal predictor performs reasonably well on the USPS data set. Therefore, it is natural to expect that the nearest neighbor(s) approach will also perform well in the problem of testing exchangeability. This is what we will use in our experiments, although in

principle almost any prediction algorithm can be adapted to testing exchangeability. The conformal transducer (deterministic or smoothed) corresponding to the nonconformity measure (3.1) (on p. 54) will be called the *NN transducer*.

When applied to the smoothed NN transducer, the family of power martingales (*NN power martingales*) might at first not look promising (Fig. 7.1), but if we concentrate on a narrower range of ϵ (Fig. 7.2), it becomes clear that the final values for some ϵ are very large.

The simple mixture of NN power martingales (which will also be referred to as the *NN SM martingale*) usually ends up with more than 10^{10} ; a typical trajectory is shown in Fig. 7.3. It is clear from this figure that the difference between the training and test sets is not the only anomaly in the USPS data set: the rapid growth of the NN SM martingale starts already on the training set.

Figure 7.3, as well as Figs. 7.5, 7.6, 7.8 referred to below, are affected by statistical variation (since the outcome depends on the random numbers τ_i actually generated), so the precise values given in the captions to those figures should not be taken too seriously. (As always in this book, we report results obtained by setting the initial state of the MATLAB pseudorandom number generator to 0.)

If p_n are output by the deterministic NN transducer, we refer to (7.5) as the *NN SM supermartingale*. As Fig. 7.4 shows, the growth rate of the latter is slightly slower than that of its randomized counterpart.

The result for a randomly permuted USPS data set is shown in Fig. 7.5. A low final value (about 1%) results from the NN SM martingale's futile attempts to gamble against an exchangeable sequence; to make possible spectacular gains against highly non-exchangeable sequences such as the USPS data set, it has to underperform against truly exchangeable sequences.

Tracking the best power martingale

The simple mixture of the previous subsection has a modest goal; the best it can do is to approximate the performance of the best power martingale. In this section we will see that it is possible to “track” the best power martingale, so that the resulting performance considerably exceeds that of the best “static” martingale (7.4).

We first generalize (7.4) as follows: for each $\epsilon = \epsilon_1 \epsilon_2 \dots \in [0, 1]^\infty$, we set

$$M_n^{(\epsilon)} := \prod_{i=1}^n (\epsilon_i p_i^{\epsilon_i - 1}) . \quad (7.6)$$

For any probability distribution μ on $[0, 1]^\infty$, define

$$M_n := \int_{[0,1]^\infty} M_n^{(\epsilon)} \mu(d\epsilon) . \quad (7.7)$$

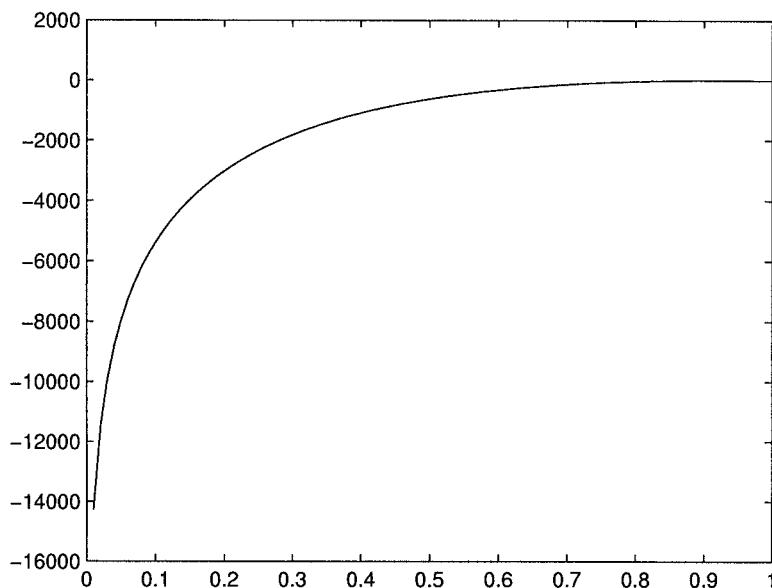


Fig. 7.1. The final values $\log_{10} M_{9298}^{(\epsilon)}$, on the logarithmic (base 10) scale, attained by the NN power martingales $M_n^{(\epsilon)}$, $0 \leq \epsilon \leq 1$, on the USPS data set

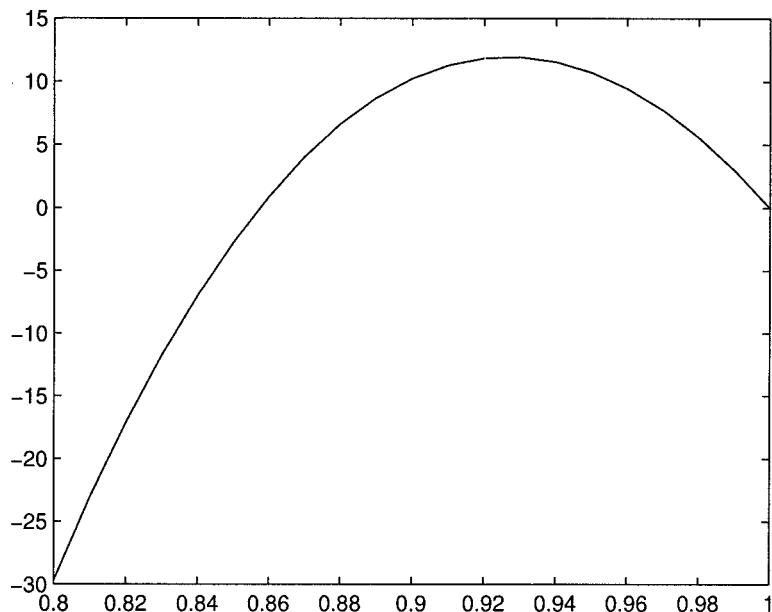


Fig. 7.2. The final values given in Fig. 7.1 for a narrower range of the parameter ϵ

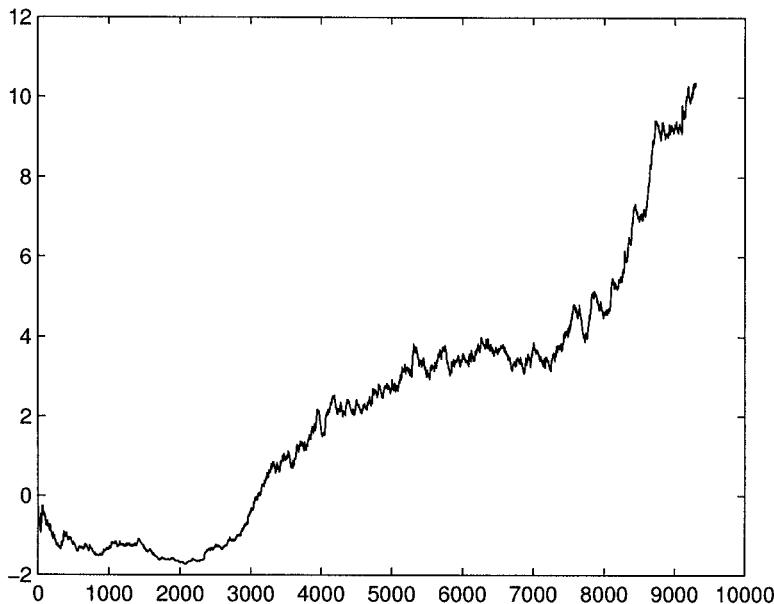


Fig. 7.3. On-line performance of the NN SM martingale on the USPS data set. The growth is shown on the logarithmic (base 10) scale: $\log M_n$ is plotted against n . The final value attained is 2.18×10^{10}

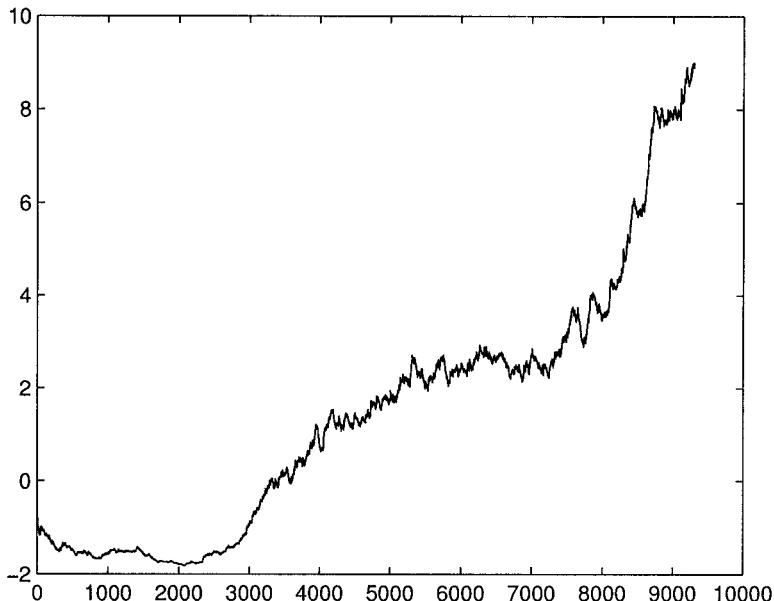


Fig. 7.4. On-line performance of the NN SM supermartingale on the USPS data set. The final value is 9.13×10^8

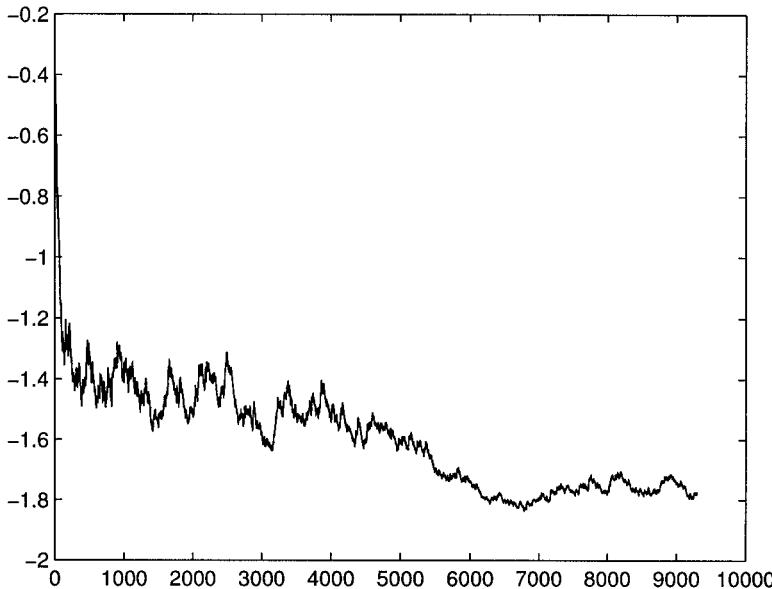


Fig. 7.5. On-line performance of the NN SM martingale on a randomly permuted USPS data set. The final value is 0.0117

It is convenient to specify μ in terms of the distribution of the coordinate random variables ϵ_n (but of course, since we integrate over μ , this does not involve any extra randomization; in particular, the mixture (7.7) is deterministic if p_n are generated by a deterministic conformal transducer). One possible μ is generated by the following *Sleepy Jumper* automaton. The states of Sleepy Jumper are elements of the Cartesian product $\{\text{awake}, \text{asleep}\} \times [0, 1]$. Sleepy Jumper starts from the state $(\text{asleep}, 1)$; when he is in a state (s, ϵ) , his transition function prescribes that:

- if $s = \text{asleep}$, he moves to the state (awake, ϵ) (“wakes up”) with probability R ($R \in [0, 1]$ is one of two parameters of the automaton) and stays in the state $(\text{asleep}, \epsilon)$ with probability $1 - R$;
- if $s = \text{awake}$, he moves to the state $(\bar{s}, \bar{\epsilon})$, where $\bar{\epsilon}$ and \bar{s} are generated independently as follows: $\bar{\epsilon} = \epsilon$ with probability $1 - J$ ($J \in [0, 1]$, the “probability of jumping”, is the other parameter) and $\bar{\epsilon}$ is chosen randomly from \mathbf{U} with probability J ; $\bar{s} = \text{awake}$ with probability $1 - R$ and $\bar{s} = \text{asleep}$ with probability R .

The output of the Sleepy Jumper automaton starting from $(s_1, \tilde{\epsilon}_1) = (\text{asleep}, 1)$ and further moving through the states $(s_2, \tilde{\epsilon}_2), (s_3, \tilde{\epsilon}_3), \dots$ is the sequence $\epsilon_1, \epsilon_2, \dots$, where

$$\epsilon_n := \begin{cases} \tilde{\epsilon}_n & \text{if } s_n = \text{awake} \\ 1 & \text{otherwise.} \end{cases}$$

The probability distribution μ of $\epsilon_1, \epsilon_2, \dots$ generated in this way defines, by (7.7), a randomized exchangeability martingale (or exchangeability supermartingale), which we call the *Sleepy Jumper martingale* (resp. *Sleepy Jumper supermartingale*). If p_n are produced by the NN transducer (smoothed or deterministic, as appropriate), we refer to the Sleepy Jumper martingale as the *NN SJ martingale* and we refer to the Sleepy Jumper supermartingale as the *NN SJ supermartingale*.

Figures 7.6 and 7.7 show the performance of the NN SJ martingale and supermartingale for parameters $R = 0.01$ and $J = 0.001$. When applied to the randomly permuted USPS data set, the NN SJ martingale's performance is as shown in Fig. 7.8. One way to improve the performance against an exchangeable data set is to decrease the jumping rate: if $J = 0.0001$, we obtain a much better performance (Fig. 7.9), even for an NN SJ supermartingale. It is easy to see the cause of the improvement: when $J = 0.0001$, the μ -measure of supermartingales (7.6) that make no jumps on the USPS data set (or any other data set of the same size) will be at least $0.9999^{9298} > e^{-1}$. The performance on the original USPS data set deteriorates (Fig. 7.10) but not drastically.

Of course, there are other ideas that can be used when combining (7.6); e.g., it would be natural to allow ϵ not only to make occasional random jumps but also to drift slowly.

Remark The approach of this section is reminiscent of “tracking the best expert” in the theory of prediction with expert advice. A general “Aggregating Algorithm” (AA) for merging experts was introduced in Vovk 1990; in the context of this section, the experts are the power martingales and the mixing operation (7.5) plays the role of (and is a special case of) the AA. Herbster and Warmuth (1998) showed how to extend the AA to “track the best expert”, to try and outperform even the best static expert. Vovk (1999) noticed that Herbster and Warmuth’s algorithm ((7.7) in the present context) is in fact a special case of the AA, when it is applied not to the original experts (in our case, (7.4)) but to “superexperts” (in our case, (7.6)).

7.2 Low-dimensional dynamic models

In Chaps. 2–6 we considered an oversimplified stochastically static picture of exchangeable environment. This picture is perhaps more useful as a building block for modeling reality rather than as a potential model.

In traditional statistics the standard building block for statistical models is *random noise* (such as the independent N_{0,σ^2} random variables ξ_i in (2.37) on p. 35; as we explain in the next chapter, this model does not require the assumption that the x_i are exchangeable). Random noise is stochastically static, but can be used as a component of dynamic models (such as (2.37) with, say, $x_i := (1, i, i^2)'$; the components of w are the dynamic parameters of this model). In typical cases we have a finite-dimensional structure combined

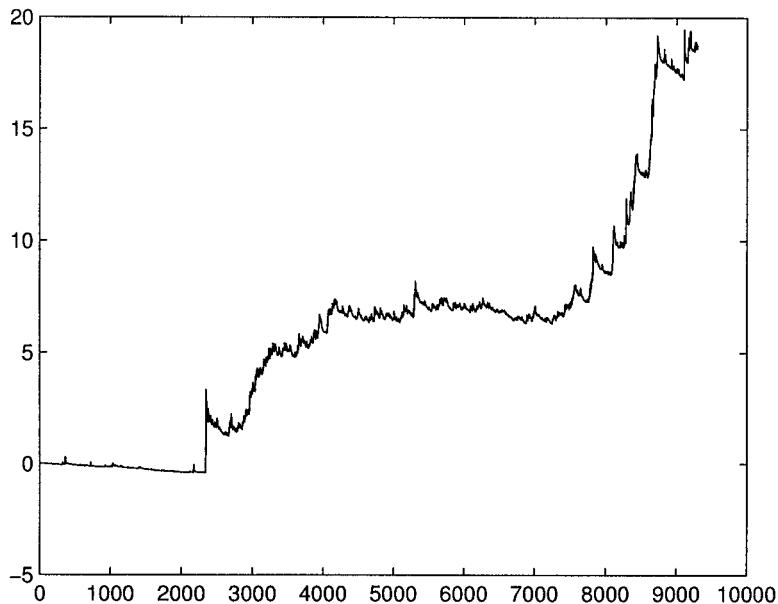


Fig. 7.6. On-line performance of the NN SJ martingale with parameters $(R, J) = (1\%, 1\%)$ on the USPS data set. The final value is 4.71×10^{18}

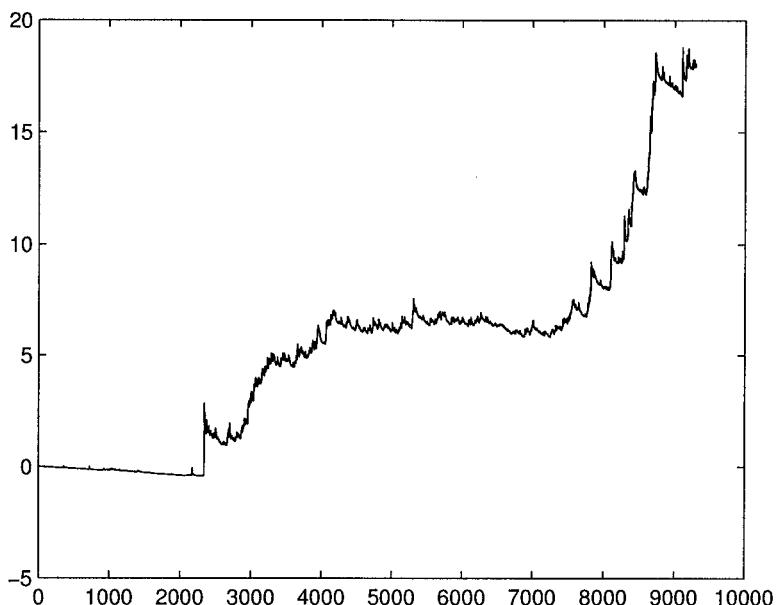


Fig. 7.7. On-line performance of the NN SJ supermartingale with parameters $(1\%, 1\%)$ on the USPS data set. The final value is 1.01×10^{18}

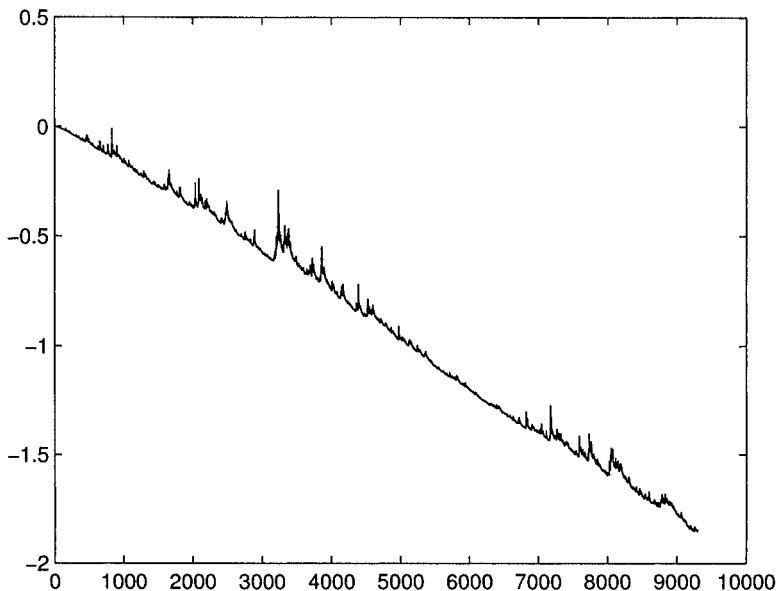


Fig. 7.8. On-line performance of the NN SJ martingale with parameters $(1\%, 1\%)$ on the randomly permuted USPS data set. The final value is 0.0142

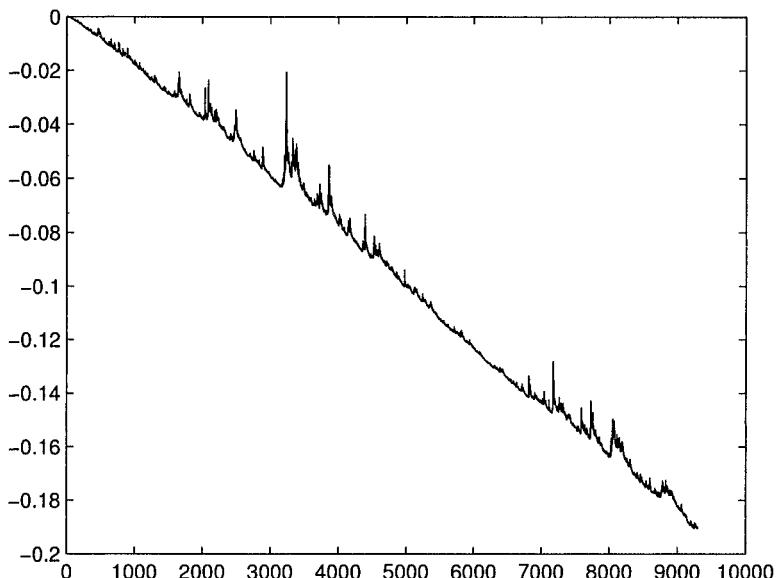


Fig. 7.9. On-line performance of the NN SJ supermartingale with parameters $(1\%, 1\%)$ on the randomly permuted USPS data set. The final value is 0.646

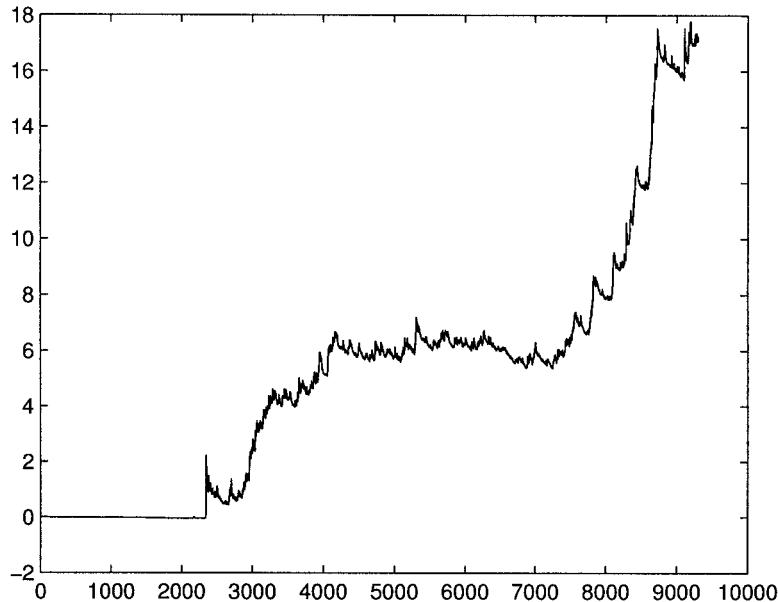


Fig. 7.10. On-line performance of the NN SJ supermartingale with parameters $(1\%, 1\%_{\text{oo}})$ on the USPS data set. The final value is 1.48×10^{17}

with a low-dimensional noise (depending on just one parameter, σ , in the case of (2.37)).

The theory developed in the previous chapters and in §7.1 allows us to replace the low-dimensional noise with a sequence of random elements about which we assume exchangeability but nothing else. We assume the existence of a “detrending transformation” which maps examples (x_i, y_i) into an exchangeable sequence (in the case of (2.37), such a detrending transformation is $(x_i, y_i) \mapsto (y_i - w \cdot x_i)$). The detrending transformation is assumed to be known except for the value of a parameter $\theta \in \Theta$. Using the methods of the previous section, we can get rid of the values of θ that do not lead to an exchangeable sequence; if the set of remaining θ is small enough, we can use the methods of conformal prediction developed in Chaps. 2–4.

As a simple example, imagine a data set containing information about house sales, where each example has the agreed house price as the label and some characteristics of the house and the area as the object (for examples of such characteristics, see the description of the Boston Housing data set in Appendix B). If the house prices have been collected over a long period of time with a constant but unknown inflation θ , we cannot assume that the data set itself is exchangeable, but the assumption of exchangeability might become more realistic after each house price y_i is multiplied by $e^{-\theta t_i}$, t_i being the time of the transaction. Replacing each label y_i with $e^{-\theta t_i} y_i$ is then a suitable detrending transformation. In this example we have a low-

dimensional dynamics (depending on just one parameter, θ) superimposed on a high-dimensional random core (the exchangeable sequence of detrended house prices).

Of course, the relevant notion of validity for dynamic models will be different from the notion of validity used in the previous chapters. As new examples are processed, we reject more and more $\theta \in \Theta$ as unlikely, and the prediction sets output by our prediction algorithms are required to be valid only with respect to the remaining θ .

Formally, a dynamic model based on exchangeability is specified by a *parametric detrending transformation*, which is a sequence of measurable functions

$$F_n : \Theta \times \mathbf{Z}^n \rightarrow \mathbf{Z}' ,$$

where $n = 1, 2, \dots$ and \mathbf{Z}' is a measurable space called the *detrended example space* (usually $\mathbf{Z}' = \mathbf{Z}$). We say that a probability distribution P on \mathbf{Z}^∞ *agrees* with the parametric detrending transformation if there is a parameter value $\theta \in \Theta$ such that the random sequence

$$F_1(\theta, z_1), F_2(\theta, z_1, z_2), \dots ,$$

where z_1, z_2, \dots are generated from P , is exchangeable.

Once we know how to produce valid (exactly or conservatively) predictions under the assumption of exchangeability and we know how to test exchangeability, it is easy to produce predictions in the dynamic model that are conservatively valid in a natural sense. For simplicity of notation we only consider deterministic confidence predictors.

Fix positive constants ϵ and δ , a parametric detrending transformation (F_n) , as above, a nonnegative exchangeability martingale S for the example space \mathbf{Z}' satisfying $S_0 = 1$, and, for each $\theta \in \Theta$, a nonconformity measure $(A_n^{(\theta)})$ for the example space \mathbf{Z}' . The *dynamic conformal predictor* determined by $\epsilon, \delta, (F_n)$, S , and $(A_n^{(\theta)})$, is defined by the equation

$$\Gamma^{\epsilon, \delta}(x_1, y_1, \dots, x_n) := \bigcup_{\theta \in \Theta} \Gamma_\theta^{\epsilon, \delta}(x_1, y_1, \dots, x_n) ,$$

where, for any $y \in \mathbf{Y}$ and $\theta \in \Theta$,

- y is not included in

$$\Gamma_\theta^{\epsilon, \delta}(x_1, y_1, \dots, x_n) \tag{7.8}$$

if

$$S_n(z_1^\theta, \dots, z_{n-1}^\theta, z_n^{\theta, y}) \geq 1/\delta ,$$

where

$$z_i^\theta := F_i(\theta, x_1, y_1, \dots, x_i, y_i), \quad i = 1, \dots, n-1,$$

$$z_n^{\theta, y} := F_n(\theta, x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, y) ;$$

- y is not included in (7.8) if

$$\frac{|\{i = 1, \dots, n : \alpha_i^{\theta,y} \geq \alpha_n^{\theta,y}\}|}{n} \leq \epsilon, \quad (7.9)$$

where

$$\begin{aligned} \alpha_i^{\theta,y} &:= A_n^{(\theta)}(z_1^\theta, \dots, z_{i-1}^\theta, z_{i+1}^\theta, \dots, z_{n-1}^\theta, z_n^{\theta,y}, z_i^\theta), \\ i &= 1, \dots, n-1, \\ \alpha_n^{\theta,y} &:= A_n^{(\theta)}(z_1^\theta, \dots, z_{n-1}^\theta, z_n^{\theta,y}). \end{aligned} \quad (7.10)$$

(cf. (2.19) on p. 26);

- y is included in (7.8) otherwise.

We will use the notation $\text{err}_n^{\epsilon,\delta}(\Gamma)$ for the indicator of error at trial n . To simplify the statement of the following proposition, we will consider the probability space that contains not only the sequence of examples z_1, z_2, \dots but also a sequence of random numbers τ_1, τ_2, \dots distributed as \mathbf{U}^∞ and independent of z_1, z_2, \dots .

Proposition 7.1. *Each dynamic conformal predictor Γ determined by $\epsilon, \delta, (F_n)$, S , and $(A_n^{(\theta)})$, as defined above, is conservative in the following sense. Suppose the data sequence z_1, z_2, \dots is generated by a probability distribution P that agrees with the parametric detrending transformation (F_n) . There exists a sequence of independent Bernoulli random variables ξ_n with parameter ϵ such that*

$$\forall n : \text{err}_n^{\epsilon,\delta}(\Gamma) \leq \xi_n$$

outside an event of P -probability δ .

Proof. Fix any parameter value θ such that $z_1^\theta, z_2^\theta, \dots$, where

$$z_i^\theta := F_i(\theta, x_1, y_1, \dots, x_i, y_i),$$

is exchangeable. Define ξ_n as the indicator of the event

$$\frac{|\{i = 1, \dots, n : \alpha_i^{\theta,y} > \alpha_n^{\theta,y}\}| + \tau_n |\{i = 1, \dots, n : \alpha_i^{\theta,y} = \alpha_n^{\theta,y}\}|}{n} \leq \epsilon, \quad (7.11)$$

where $\alpha_i^{\theta,y}$ are defined by (7.10) (event (7.11) is the smoothed counterpart of (7.9)) and apply Doob's inequality (§A.6). \square

7.3 Islands of randomness

In this section we assume that instead of a comprehensive theory explaining all observations we only have a patchwork of theories each explaining only a relatively small piece of the observed data sequence. We will discuss only

the simplest case where each of the “local theories” is just the hypothesis of randomness applied to a subsequence of the full data sequence. We first prove a simple mathematical result about the conservative validity of conformal predictors in this case, and then briefly discuss how this result applies to the prediction of Markov sequences (we will return to the topic of Markov sequences in the next chapter).

A sufficient condition for asymptotic validity

The random subsequences about which the assumption of randomness is made will be chosen in a “predictable” manner, in the spirit of von Mises’s subsequence selection rules (see, e.g., Shafer and Vovk 2001, §2.3). Formal definitions will use filtrations to formalize the intuitive notion of information available at different times (cf. §A.6).

The observed examples $z_n = (x_n, y_n)$, $n = 1, 2, \dots$, are random elements, taking values in the example space $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ (with x_n and y_n taking values in \mathbf{X} and \mathbf{Y} , respectively), defined on an underlying probability space. Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \dots$ be a filtration on this probability space such that the sequence x_n is predictable and the sequence y_n is adapted, w.r. to this filtration. (In the sequel, “w.r. to this filtration” will be omitted.) Intuitively, \mathcal{F}_{n-1} is the information available when making prediction for the label y_n ; the requirement that the sequence \mathcal{F}_n should be increasing means that nothing is ever forgotten, and the requirement that x_n be predictable and y_n be adapted means that z_1, \dots, z_{n-1}, x_n are known when making the prediction for y_n . When we are interested in randomized predictors, \mathbf{X} is interpreted as the set of extended objects.

Let K be a finite or countable set (indexing the subsequences of z_1, z_2, \dots we are going to consider) and, for each $k \in K$, let ν_1^k, ν_2^k, \dots be a predictable sequence of binary random variables, taking values in $\{0, 1\}$. Intuitively, $\nu_n^k = 1$ means that we include the example z_n in the subsequence indexed by k . We will assume that each example belongs to no more than one subsequence: $\sum_{k \in K} \nu_n^k \leq 1$, for all $n = 1, 2, \dots$.

Let z_i^k stand for the example z_n where n is defined by the requirements

$$\nu_n^k = 1 \text{ \& } \sum_{j=1}^n \nu_j^k = i,$$

if such an n exists (z_i^k is undefined if it does not exist). Therefore, z_1^k, z_2^k, \dots is the k th subsequence (infinite or finite) of the full data sequence z_1, z_2, \dots

Remark The framework of this section is closely related to the framework in which Mondrian predictors are analyzed (see Chap. 4 and, especially, the next chapter). A Mondrian taxonomy is also a way of splitting the full data sequence into subsequences; the decision whether z_n is included depends, however, on n and z_n rather than on z_1, \dots, z_{n-1}, x_n (and whatever extra information may be encoded in \mathcal{F}_{n-1}).

Let

$$I_n := \sum_{k \in K} \sum_{i=1}^n \nu_i^k$$

be the number of examples among z_1, \dots, z_n that are covered by the subsequences, and let

$$a_n := \left| \left\{ k \in K : \sum_{i=1}^n \nu_i^k > 0 \right\} \right|$$

be the number of subsequences that started before or at trial n .

We assume that the examples within each subsequence z_1^k, z_2^k, \dots are independent and identically distributed. Let Γ be a confidence predictor (such as a conformal predictor or a smoothed conformal predictor); running a different copy of Γ within each z_1^k, z_2^k, \dots , we obtain another confidence predictor $\tilde{\Gamma}$. Formally, $\tilde{\Gamma}^\epsilon(x_1, y_1, \dots, x_n)$ is defined as follows:

- if $\nu_n^k = 0$ for all $k \in K$, set $\tilde{\Gamma}^\epsilon(x_1, y_1, \dots, x_n) := \mathbf{Y}$;
- otherwise, set

$$\tilde{\Gamma}^\epsilon(x_1, y_1, \dots, x_n) := \Gamma^\epsilon(z_1^k, \dots, z_{\sum_{j=1}^n \nu_j^k - 1}^k, x_n),$$

where k is such that $\nu_n^k = 1$.

As usual, $\text{Err}_n^\epsilon(\tilde{\Gamma})$ stands for the number of errors made by the composite predictor $\tilde{\Gamma}$ at significance level ϵ up to (and including) trial n . The following proposition (proved in the next section) asserts that the composite predictor is asymptotically valid unless the number of random subsequences is very large.

Proposition 7.2. *Suppose that, in the notation introduced above, the examples within each subsequence z_1^k, z_2^k, \dots are independent and identically distributed. If a confidence predictor Γ is exact, the event*

$$\left(I_n \rightarrow \infty \& a_n = o\left(\frac{I_n}{\ln I_n}\right) \right) \implies \lim_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon(\tilde{\Gamma})}{I_n} = \epsilon \quad (7.12)$$

has probability one for each $\epsilon \in (0, 1)$. If Γ is conservative but not necessarily exact, (7.12) will continue to hold if “ \lim ” is replaced by “ \limsup ” and “ $= \epsilon$ ” is replaced by “ $\leq \epsilon$ ”.

Markov sequences

One of the simplest applications of Proposition 7.2 is to sequences generated by Markov chains. For the definition of Markov chains, see, e.g., Shiryaev 1996 (§I.12 and Chap. VIII).

Consider a Markov chain with a finite set of states \mathbf{Z} and take the observed sequence of its states as the sequence of examples z_1, z_2, \dots . Set $K := \mathbf{Z}$,

$$\nu_n^k := \begin{cases} 1 & \text{if } k = z_{n-1} \\ 0 & \text{otherwise,} \end{cases}$$

for $n = 2, 3, \dots$, and $\nu_1^k := 0$ for all k . In other words, the k th random subsequence consists of the examples z_n coming after $z_{n-1} = k$; z_1 is the only example that does not belong to any random subsequence. Since $I_n = n - 1$ and $a_n \leq |\mathbf{Z}|$, Proposition 7.2 is applicable, and so $\tilde{\Gamma}$ will be asymptotically exact for any smoothed conformal predictor Γ ; for deterministic conformal predictors Γ we will have asymptotic conservative validity.

Prediction of Markov chains will also be considered, from a different point of view, in Chap. 8. There we will construct valid, not only asymptotically valid, confidence predictors for Markov chains.

7.4 Proof of Proposition 7.2

A natural idea is to use Hoeffding's inequality, but the difficulty (familiar from the previous chapter) is that the number of examples among z_1, \dots, z_n in the k th subsequence will be random, and the inequality requires the number of examples to be deterministic. We will have to use the fact that (A.12) (on p. 288) is a supermartingale directly.

Fix a significance level $\epsilon \in (0, 1)$; we will omit the index ϵ . For each k , let

$$N_n^k := \sum_{i=1}^n \nu_i^k$$

be the number of examples z_1, \dots, z_n included in the k th subsequence, and let

$$\text{Err}_n^k := \sum_{i=1}^n \text{Err}_i(\tilde{\Gamma}) \nu_i^k$$

be the number of errors made by $\tilde{\Gamma}$ on those of the first n examples z_1, \dots, z_n that belong to the k th subsequence. (Less formally, Err_n^k is the number of errors made by the k th copy of Γ on the examples z_1, \dots, z_n .)

Since a_n depends on n only through I_n , we can introduce a (random) function A that maps I_n to a_n ; we will write A_I for $A(I)$. Without loss of generality we assume that $I_n = n$, omitting the examples z_n that do not belong to any subsequence.

We first consider the case where the confidence predictor Γ is conservative. As explained in §A.7 (see (A.12) on p. 288), the sequence of random variables

$$M_n^{\kappa, k} := \exp \left(\kappa S_n^k - \frac{\kappa^2}{8} N_n^k \right), \quad n = 1, 2, \dots,$$

where κ is a positive rational constant and

$$S_n^k := \text{Err}_n^k - \epsilon N_n^k,$$

is a supermartingale. Since their combination

$$M := \sum_{\kappa \in \mathbb{Q} \cap (0, \infty)} \sum_{i=1}^{\infty} w(\kappa) i^{-2} M^{\kappa, k_i},$$

where k_i are the indices of the subsequences in the order of their first appearance¹ and $w(\kappa)$ are positive weights such that

$$\sum_{\kappa \in \mathbb{Q} \cap (0, \infty)} \sum_{i=1}^{\infty} w(\kappa) i^{-2} = 1,$$

is also a supermartingale, Doob's inequality (see §A.6) implies that the random variable $C := \sup_n |M_n|$ is finite with probability one. Therefore, with probability one there exists a $C < \infty$ such that

$$\exp \left(\kappa S_n^k - \frac{\kappa^2}{8} N_n^k \right) \leq C A_n^2 / w(\kappa)$$

for all n and k . Taking the logarithm of both sides and summing over the A_n indices k of the subsequences that are present in z_1, \dots, z_n , we obtain

$$\kappa S_n - \frac{\kappa^2}{8} n \leq A_n (2 \ln A_n + d(\kappa)),$$

where $d(\kappa) := \ln(C/w(\kappa))$ and $S_n := \text{Err}_n(\tilde{\Gamma}) - \epsilon n$. Therefore,

$$\frac{\text{Err}_n(\tilde{\Gamma})}{n} \leq \epsilon + \frac{\kappa}{8} + \frac{A_n (2 \ln A_n + d(\kappa))}{\kappa n}.$$

Letting $n \rightarrow \infty$ and then $\kappa \rightarrow 0$, we obtain that

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n(\tilde{\Gamma})}{n} \leq \epsilon \tag{7.13}$$

provided $A_n \ln A_n = o(n)$. The last condition follows from $A_n = o(n/\ln n)$: indeed, if $A_n < \delta n/\ln n$ for some $\delta > 0$ from some n on, then (using the fact that $a \ln a$ is a strictly increasing function of a for large a)

$$A_n \ln A_n < \delta \frac{n}{\ln n} \ln \left(\delta \frac{n}{\ln n} \right) < \delta n$$

from some n on. This proves the second statement of the proposition.

¹The formal inductive definition is: k_1 is defined by the requirement $\nu_1^{k_1} = 1$; k_i , $i = 2, 3, \dots$, is defined by the requirements that $k_i \notin \{k_1, \dots, k_{i-1}\}$ and there exists an n such that $\nu_n^{k_i} = 1$ and $\nu_j^k = 0$ for all $j = 1, \dots, n-1$ and all $k \notin \{k_1, \dots, k_{i-1}\}$.

Let us now assume that Γ is not only conservative but also exact. Taking κ negative, we in the same way prove that

$$\liminf_{n \rightarrow \infty} \frac{\text{Err}_n(\tilde{\Gamma})}{n} \geq \epsilon \quad (7.14)$$

provided $A_n = o(n/\ln n)$. Combining (7.13) and (7.14), we obtain (7.12).

In conclusion, we will explicitly give a definition that is not quite standard but was used in stating the proposition. The random elements, taking values in a measurable space \mathbf{Z} , in a sequence ζ_1, ζ_2, \dots of random length (possibly infinite) are *independent and identically distributed* if there is a probability distribution Q on \mathbf{Z} such that, for each $n = 1, 2, \dots$, the conditional distribution of ζ_n given that ζ_1, \dots, ζ_n exist and given the values of $\zeta_1, \dots, \zeta_{n-1}$ is Q .

7.5 Bibliographical remarks

In §7.1 we follow mainly Vovk et al. 2003b. Before that paper, it was not even clear that nontrivial exchangeability supermartingales exist; we saw that they not only exist, but can attain huge final values on a benchmark (USPS) data set starting from 1 and never risking bankruptcy.

Definition (7.4) is based on the procedure suggested in Vovk 1993 (§9).

The definition of dynamic models based on exchangeability given in §7.2 is motivated by Barnard's (1977) pivotal inference (the idea of using pivots was suggested to him by Fisher; see DeGroot 1988, p. 202). For references to related literature, see Dawid and Stone 1982 (especially Fraser's comment).

The proof in §7.4 is similar to the many martingale proofs in Shafer and Vovk 2001.

On-line compression modeling I: conformal prediction

We know that each conformal predictor is automatically valid when used in the on-line mode and provided that the data sequence is generated by an exchangeable distribution. In this chapter we state a more general result replacing the assumption of exchangeability of the data-generating distribution by the assumption that the data agrees with a given “on-line compression model”; the exchangeability model is just one of many interesting models of this type. This chapter’s result is a step towards implementation of Kolmogorov’s program for applications of probability; in particular, the concept of on-line compression model is an on-line version of the concept considered by Kolmogorov (and is closely connected to Martin-Löf’s repetitive structures and Freedman’s summarizing statistics).

In §8.1 we define the on-line compression models, which include, besides the exchangeability model, the Gaussian model, the Markov model, and many other interesting models. An on-line compression model (OCM) is an automaton (usually infinite) for summarizing statistical information efficiently. It is usually impossible to restore the statistical information from the OCM’s summary (so OCM performs lossy compression), but it can be argued that the only information lost is noise, since one of our requirements is that the summary should be a “sufficient statistic”. In §8.2 we construct conformal transducers for an arbitrary OCM and state a simple theorem (proved in §8.7) showing that the confidence information provided by conformal transducers is valid; this theorem generalizes the validity result of Chap. 2. We then briefly remind the reader how conformal transducers are used for confidence prediction. In §8.3 we describe an alternative language for on-line compression modeling; it is based on Martin-Löf’s notion of repetitive structure and is very convenient when specific models are discussed. In the following three sections, §8.4–8.6, we consider three interesting examples of on-line compression models: exchangeability, Gaussian and Markov models. In §8.8 we discuss the origins of the idea of on-line compression modeling and of specific on-line compression models.

8.1 On-line compression models

We are interested in making predictions about a sequence of examples z_1, z_2, \dots output by Reality. Typically we will want to say something about example z_n , $n = 1, 2, \dots$, given the previous examples z_1, \dots, z_{n-1} . In this section we will discuss an assumption that we might be willing to make about the examples, and in the next section the actual prediction algorithms.

An *on-line compression model* (OCM) is a 5-tuple

$$M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n)) ,$$

where:

1. Σ is a measurable space called the *summary space*; its elements are called *summaries*; $\square \in \Sigma$ is a summary called the *empty summary*;
2. \mathbf{Z} is a measurable space from which the examples z_i are drawn;
3. F_n , $n = 1, 2, \dots$, are measurable functions of the type $\Sigma \times \mathbf{Z} \rightarrow \Sigma$ called *forward functions*;
4. B_n , $n = 1, 2, \dots$, are Markov kernels (see §A.4) of the type $\Sigma \hookrightarrow \Sigma \times \mathbf{Z}$ called *backward kernels*; it is required that B_n be an inverse to F_n in the sense that

$$B_n(F_n^{-1}(\sigma) | \sigma) = 1$$

for each $\sigma \in F_n(\Sigma \times \mathbf{Z})$.

Next we explain briefly the intuition behind this formal definition and introduce some further notation.

An OCM is a way of summarizing statistical information. At the beginning we do not have any information, which is represented by the empty summary $\sigma_0 := \square$. When the first example z_1 arrives, we update our summary to $\sigma_1 := F_1(\sigma_0, z_1)$, etc.; when example z_n arrives, we update the summary to $\sigma_n := F_n(\sigma_{n-1}, z_n)$. This process is represented in Fig. 8.1. Let t_n be the *nth statistic* in the OCM, which maps the sequence of the first n examples z_1, \dots, z_n to σ_n :

$$\begin{aligned} t_1(z_1) &:= F_1(\sigma_0, z_1); \\ t_n(z_1, \dots, z_n) &:= F_n(t_{n-1}(z_1, \dots, z_{n-1}), z_n), \quad n = 2, 3, \dots . \end{aligned} \tag{8.1}$$

The value $t_n(z_1, \dots, z_n)$ is a summary of the full data sequence z_1, \dots, z_n available at the end of trial n ; our definition requires that the summaries should be computable on-line: the function F_n updates σ_{n-1} to σ_n .

Condition 3 in the definition of OCM reflects its on-line character, as explained in the previous paragraph. We want, however, the system of summarizing statistical information represented by the OCM to be accurate, so that no useful information is lost. This is reflected in Condition 4: the distribution P_n of the more detailed description (σ_{n-1}, z_n) given the less detailed σ_n is known, and so (σ_{n-1}, z_n) does not carry any additional information

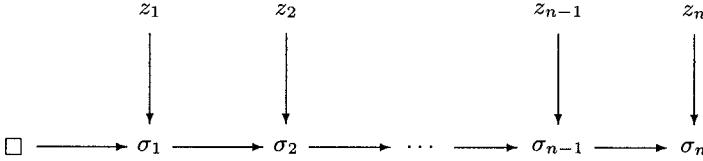


Fig. 8.1. Using the forward functions F_n to compute σ_n from z_1, \dots, z_n

about the distribution generating the examples z_1, z_2, \dots ; in other words, σ_n contains the same useful information as (σ_{n-1}, z_n) , and the extra information in (σ_{n-1}, z_n) is noise. This intuition would be captured in statistical terminology (see, e.g., Cox and Hinkley 1974, §2.2) by saying that σ_n is a “sufficient statistic” of (σ_{n-1}, z_n) and, eventually, of z_1, \dots, z_n (although this expression does not have a formal meaning in our present context, since we do not have a full statistical model $(P_\theta : \theta \in \Theta)$ at this point).

Analogously to Fig. 8.1, we can find the distribution of the data sequence z_1, \dots, z_n from σ_n (see Fig. 8.2): given σ_n , we generate the pair (σ_{n-1}, z_n) from the distribution $B_n(\sigma_n)$, then we generate (σ_{n-2}, z_{n-1}) from $B_{n-1}(\sigma_{n-1})$, etc. Formally, using the Markov kernels $B_n(d\sigma_{n-1}, dz_n | \sigma_n)$, we can define the conditional distribution P_n of z_1, \dots, z_n given σ_n by the requirement that, for all bounded measurable functions $f : \mathbf{Z}^n \rightarrow \mathbb{R}$,

$$\begin{aligned} \int f(z_1, \dots, z_n) P_n(dz_1, \dots, dz_n | \sigma_n) := \\ \int \cdots \int f(z_1, \dots, z_n) B_1(dz_1 | \sigma_1) B_2(dz_2 | \sigma_2) \cdots \\ B_{n-1}(d\sigma_{n-2}, dz_{n-1} | \sigma_{n-1}) B_n(d\sigma_{n-1}, dz_n | \sigma_n); \end{aligned} \quad (8.2)$$

the existence of such a probability distribution P_n immediately follows from the Stone–Daniell theorem (see, e.g., Dudley 2002, Theorem 4.5.2). A shorter way to write (8.2) is

$$\begin{aligned} P_n(dz_1, \dots, dz_n | \sigma_n) := B_1(dz_1 | \sigma_1) B_2(d\sigma_1, dz_2 | \sigma_2) \cdots \\ B_{n-1}(d\sigma_{n-2}, dz_{n-1} | \sigma_{n-1}) B_n(d\sigma_{n-1}, dz_n | \sigma_n). \end{aligned}$$

We say that a probability distribution P on \mathbf{Z}^∞ *agrees* with the OCM $(\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ if, for each n and each event $A \subseteq \Sigma \times \mathbf{Z}$, $B_n(A | \sigma)$ is a version of the conditional probability, w.r. to P , that $(t_{n-1}(z_1, \dots, z_{n-1}), z_n) \in A$ given $t_n(z_1, \dots, z_n) = \sigma$ and given the values of z_{n+1}, z_{n+2}, \dots .

We will write Σ_n for $t_n(\mathbf{Z}^n)$, $n \geq 1$, and Σ_0 for $\{\square\}$; the elements of Σ_n will be called *n-summaries*. Notice that F_n maps $\Sigma_{n-1} \times \mathbf{Z}$ to Σ_n , and the probability distribution $B_n(\sigma)$ is concentrated on $\Sigma_{n-1} \times \mathbf{Z}$ for all $\sigma \in \Sigma_n$.

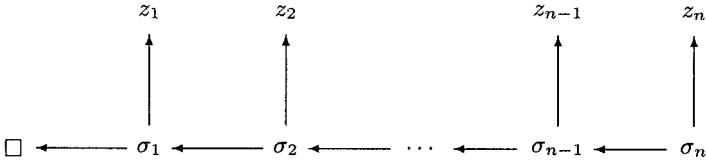


Fig. 8.2. Using the backward kernels B_n to extract the distribution of z_1, \dots, z_n from σ_n

All our definitions and results will involve only these restrictions: of F_n to $\Sigma_{n-1} \times \mathbf{Z}$ and of B_n to Σ_n .

It is sometimes convenient to use a slightly modified definition of an on-line compression model removing the “unused” part of Σ . If $(\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ is an on-line compression model, the corresponding *reduced on-line compression model* is $((\Sigma_n), \square, \mathbf{Z}, (F'_n), (B'_n))$ (n ranging over \mathbb{N}), where

$$F'_n := F_n|_{\Sigma_{n-1} \times \mathbf{Z}}, \quad B'_n := B_n|_{\Sigma_n},$$

$f|_E$ standing for the restriction of f to E .

8.2 Conformal transducers and validity of OCM

In Chap. 2, any function f of the type $(\mathbf{Z} \times [0, 1])^* \rightarrow [0, 1]$ was called a *randomized transducer*; it is regarded as mapping each input sequence $(z_1, \tau_1, z_2, \tau_2, \dots)$ in $(\mathbf{Z} \times [0, 1])^\infty$ into the output sequence of *p-values* (p_1, p_2, \dots) defined by $p_n := f(z_1, \tau_1, \dots, z_n, \tau_n)$, $n = 1, 2, \dots$. We say that the randomized transducer f is *exactly valid* w.r. to an OCM M if the output p-values $p_1 p_2 \dots$ are always distributed according to the uniform distribution \mathbf{U}^∞ on $[0, 1]^\infty$, provided the input examples $z_1 z_2 \dots$ are generated by a probability distribution that agrees with M and $\tau_1 \tau_2 \dots$ are generated, independently of $z_1 z_2 \dots$, from \mathbf{U}^∞ . If we drop the dependence on the random numbers τ_n , we obtain the notion of *deterministic transducer*.

Any sequence of measurable functions $A_n : \Sigma_{n-1} \times \mathbf{Z} \rightarrow \mathbb{R}$, $n = 1, 2, \dots$, is called a *nonconformity measure* w.r. to the OCM $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$. The *conformal transducer* determined by (A_n) is the deterministic transducer where p_n are defined as

$$p_n := B_n(\{(\sigma, z) \in \Sigma_{n-1} \times \mathbf{Z} : A_n(\sigma, z) \geq A_n(\sigma_{n-1}, z_n)\} | \sigma_n) \quad (8.3)$$

and

$$\sigma_n := t_n(z_1, \dots, z_n), \quad \sigma_{n-1} := t_{n-1}(z_1, \dots, z_{n-1}).$$

The randomized version, called the *smoothed conformal transducer* determined by (A_n) , is obtained by replacing (8.3) with

$$\begin{aligned} p_n := & B_n (\{(\sigma, z) \in \Sigma_{n-1} \times \mathbf{Z} : A_n(\sigma, z) > A_n(\sigma_{n-1}, z_n)\} \mid \sigma_n) \\ & + \tau_n B_n (\{(\sigma, z) \in \Sigma_{n-1} \times \mathbf{Z} : A_n(\sigma, z) = A_n(\sigma_{n-1}, z_n)\} \mid \sigma_n). \end{aligned} \quad (8.4)$$

A *conformal transducer* in an OCM M is a conformal transducer (deterministic or smoothed) determined by some nonconformity measure w.r. to M .

Theorem 8.1. *Suppose the examples $z_n \in \mathbf{Z}$, $n = 1, 2, \dots$, are generated from a probability distribution P that agrees with an on-line compression model. Any smoothed conformal transducer in that model is exactly valid (will produce independent p-values p_n distributed uniformly in $[0, 1]$).*

As discussed in Chap. 2, conformal transducers can be used for hedged prediction; we will briefly summarize how this is done. Suppose each example z_n consists of two components, x_n (the object) and y_n (the label); at trial n we are given x_n and the goal is to predict y_n ; for simplicity, we will assume that the label space \mathbf{Y} from which the labels are drawn is finite. Suppose we are given a significance level $\epsilon > 0$ (the maximum probability of error we are prepared to tolerate). When given x_n , we can output as the prediction set $\Gamma_n^\epsilon \subseteq \mathbf{Y}$ the set of labels y such that $y_n = y$ would lead to a p-value $p_n > \epsilon$. (When a conformal transducer is applied in this mode, it is referred to as a *conformal predictor*.) If an error at trial n is defined as $y_n \notin \Gamma_n^\epsilon$, then by Theorem 8.1 errors at different trials are independent and the probability of error at each trial is ϵ , assuming the p_n are produced by a smoothed conformal transducer. In particular, such confidence predictors are asymptotically exact, in the sense that the number Err_n^ϵ of errors made in the first n trials satisfies

$$\lim_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon}{n} = \epsilon \quad \text{a.s.}$$

This implies that if the p_n are produced by a deterministic conformal transducer, we will still have the conservative version of this property,

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n^\epsilon}{n} \leq \epsilon \quad \text{a.s.}$$

Finite-horizon result

In this subsection we will state a modification of Theorem 8.1 which, although slightly less elegant than Theorem 8.1 itself, is mathematically stronger and more readily applicable (in particular, it explains why shuffling finite data sets, as described in §B.4, makes smoothed conformal transducers produce independent p-values distributed uniformly on $[0, 1]$).

Theorem 8.1 is so general that it implies almost all other validity results in this book, but its generality also gives rise to some problems. The statement of the theorem is given in terms of probability distributions P that agree with the given OCM, but even the simplest questions about the class \mathcal{P} of such P are very difficult and can at present be answered only in some special

cases (see, e.g., Lauritzen 1988; the exposition in that book is in terms of “repetitive structures”, which, as explained in the next section, provide an equivalent language for talking about on-line compression modeling). Even the question whether \mathcal{P} is non-empty is difficult (although the answer is known to be positive if \mathbf{Z} is finite; see Proposition 7.1 in Lauritzen 1988, p. 75). The question whether \mathcal{P} is nontrivial (i.e., whether $|\mathcal{P}| > 1$) is even more difficult; in statistical mechanics this question appears as the question of existence of phase transition. Moreover, the mathematical techniques used to answer these questions are heavily asymptotic and seem very remote from the practice of machine learning.

Fix a positive integer N , the *horizon*. An *on-line compression N-model* (or *N-OCM*) is defined in the same way as an ordinary OCM but with n ranging over the set $\{1, \dots, N\}$. Statistics t_n and conditional probability distributions P_n , $n = 1, \dots, N$, are defined by (8.1) and (8.2). We say that a probability distribution P on \mathbf{Z}^N *agrees* with an *N-OCM* $(\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ if, for each $n = 1, \dots, N$ and each event $A \subseteq \Sigma \times \mathbf{Z}$, $B_n(A | \sigma)$ is a version of the conditional probability, w.r. to P , that $(t_{n-1}(z_1, \dots, z_{n-1}), z_n) \in A$ given $t_n(z_1, \dots, z_n) = \sigma$ and given the values of z_{n+1}, \dots, z_N . The description of the family of all P that agree with the *N-OCM* is trivial: these are the mixtures of $P_N(\cdot | \sigma)$ over $\sigma \in \Sigma_N = t_N(\mathbf{Z}^N)$. (Indeed, each $P_N(\cdot | \sigma)$ agrees with the *N-OCM* and each P that agrees with the *N-OCM* is $\int_{\Sigma} P_N(\cdot | \sigma) P'(\mathrm{d}\sigma)$, where P' is the image of P under the mapping t_N .)

Nonconformity measures and conformal transducers for *N-OCM* are defined as before, with the only difference that n now ranges in $\{1, \dots, N\}$.

Theorem 8.2. *Let $N \in \mathbb{N}$ and the examples $z_n \in \mathbf{Z}$, $n = 1, \dots, N$, be generated from a probability distribution P on \mathbf{Z}^N that agrees with an on-line compression *N-model*. Any smoothed conformal transducer in that model is exactly valid (will produce independent p-values p_n , $n = 1, \dots, N$, distributed uniformly in $[0, 1]$).*

In particular, any smoothed conformal predictor will produce p-values distributed as \mathbf{U}^N if the examples are generated from any conditional distribution $P_N(\sigma)$, $\sigma \in \Sigma_N$, in the notation introduced above. This shows that Theorem 8.2 indeed implies Proposition 4.10 on p. 115.

It is clear that Theorem 8.2 implies Theorem 8.1: if a probability distribution P on \mathbf{Z}^∞ agrees with an OCM, the restriction of P to the first N examples will agree with the restriction of the OCM to the first N examples, and therefore, the first N p-values p_1, \dots, p_N will be distributed according to the uniform distribution on $[0, 1]^N$; now standard results (such as Williams 1991, Lemma 1.6) imply that the infinite sequence p_1, p_2, \dots has the uniform distribution on $[0, 1]^\infty$.

8.3 Repetitive structures

There are two equivalent languages for discussing on-line compression modeling: our on-line compression models and Martin-Löf's repetitive structures. The former is more convenient in the general theory, considered so far, and the latter is better suited to discussing specific models, which we do in the following sections. In this section we define repetitive structures and compare the two languages.

Let Σ and \mathbf{Z} be measurable spaces (of “summaries” and “examples”, respectively). A *repetitive structure* $(\Sigma, \mathbf{Z}, (t_n), (P_n))$ contains, additionally, the following two elements:

- a system of statistics (measurable functions) $t_n : \mathbf{Z}^n \rightarrow \Sigma$, $n = 1, 2, \dots$;
- a system of Markov kernels $P_n : \Sigma \hookrightarrow \mathbf{Z}^n$, $n = 1, 2, \dots$.

These two elements are required to satisfy the following consistency requirements:

Agreement between P_n and t_n : for each $\sigma \in t_n(\mathbf{Z}^n)$, the probability distribution $P_n(\cdot | \sigma)$ is concentrated on the set $t_n^{-1}(\sigma)$;

On-line character of t_n : for all integers $n > 1$, $t_n(z_1, \dots, z_n)$ is determined by $t_{n-1}(z_1, \dots, z_{n-1})$ and z_n , in the sense that the function t_n is measurable w.r. to the σ -algebra generated by t_{n-1} and z_n ;

Consistency of P_n : for all integers $n > 1$ and all $\sigma_n \in t_n(\mathbf{Z}^n)$, $P_{n-1}(\cdot | \sigma_{n-1})$ should be a version of the conditional distribution of z_1, \dots, z_{n-1} when z_1, \dots, z_n is generated from $P_n(dz_1, \dots, dz_n | \sigma_n)$ and it is known that $t_{n-1}(z_1, \dots, z_{n-1}) = \sigma_{n-1}$ and $z_n = z$ (σ_{n-1} ranging over $t_{n-1}(\mathbf{Z}^{n-1})$ and z over \mathbf{Z}).

The *reduced version* of a repetitive structure $(\Sigma, \mathbf{Z}, (t_n), (P_n))$ is defined to be $((\Sigma_n), \mathbf{Z}, (t'_n), (P'_n))$, where $\Sigma_n := t_n(\mathbf{Z}^n)$, t'_n is the same as t_n but of the type $\mathbf{Z}^n \rightarrow \Sigma_n$, and $P'_n := P_n|_{\Sigma_n}$.

The on-line character of t_n can be restated as follows: there exists a sequence of measurable functions $F_n : \Sigma_{n-1} \times \mathbf{Z} \rightarrow \Sigma_n$, $n = 2, 3, \dots$, such that

$$t_n(z_1, \dots, z_n) = F_n(t_{n-1}(z_1, \dots, z_{n-1}), z_n) \quad (8.5)$$

for all n and $z_1, \dots, z_n \in \mathbf{Z}$. This makes repetitive structures more similar to on-line compression models; the full equivalence is established in the following proposition.

Proposition 8.3. *If $M = ((\Sigma_n), \square, \mathbf{Z}, (F_n), (B_n))$ is a reduced on-line compression model, then $M' := ((\Sigma_n), \mathbf{Z}, (t_n), (P_n))$ (see (8.1) and (8.2)) is a reduced repetitive structure. If $M = ((\Sigma_n), \mathbf{Z}, (t_n), (P_n))$ is a reduced repetitive structure, a reduced on-line compression model $M' = ((\Sigma_n), \square, \mathbf{Z}, (F_n), (B_n))$ can be defined as follows:*

- \square is, e.g., the empty set;

- F_n are the functions from (8.5), $n = 2, 3, \dots$, and $F_1(\square, z) := t_1(z)$ for all $z \in \mathbf{Z}$;
- $B_n(d\sigma_{n-1}, dz_n | \sigma_n)$ is the image of the distribution $P_n(dz_1, \dots, dz_n | \sigma_n)$ under the mapping

$$(z_1, \dots, z_n) \mapsto (\sigma_{n-1}, z_n),$$

where $\sigma_{n-1} := t_{n-1}(z_1, \dots, z_{n-1})$.

If M is a reduced on-line compression model, $M'' = M$. If M is a reduced repetitive structure, $M'' = M$.

8.4 Exchangeability model and its modifications

In this section we discuss some familiar (although not defined formally so far) on-line compression models, viz., the exchangeability model and its modifications; in the next two sections we will consider new models, Gaussian and Markov. All specific OCMs will be defined through their statistics t_n and conditional distributions P_n (i.e., through the corresponding repetitive structure, which will be called the “repetitive-structure representation” of the OCM). For prediction, however, it will be important to move to the representation as an on-line compression model (the “on-line compression representation”, as we will say).

Exchangeability model

The *exchangeability model* has statistics

$$t_n(z_1, \dots, z_n) := \lceil z_1, \dots, z_n \rceil ;$$

given the value of the statistic, all orderings have the same probability $1/n!$. Remember that formally we define the set of bags $\lceil z_1, \dots, z_n \rceil$ of size n to be the power set \mathbf{Z}^n equipped with the σ -algebra of symmetric (i.e., invariant under permutations of components) events; the probability distribution on the orderings is given by $z_{\pi(1)}, \dots, z_{\pi(n)}$, where z_1, \dots, z_n is a fixed ordering and π is a random permutation (each permutation is chosen with probability $1/n!$).

It is easy to see what the on-line compression representation of the exchangeability model is. The function $F_n : (\sigma_{n-1}, z_n) \mapsto \sigma_n$ puts another example z_n in the bag σ_{n-1} producing a bigger bag σ_n . The probability distribution $B_n(\sigma_n)$ can be implemented as follows: draw an example z_n from the bag σ_n at random and output the pair (σ_{n-1}, z_n) , where σ_{n-1} is σ_n with z_n removed.

The set of probability distributions on \mathbf{Z}^∞ that agree with the exchangeability OCM is exactly the exchangeability statistical model (Lemma A.3 on p. 283 shows that each exchangeable probability distribution agrees with the

exchangeability OCM, and Lemma A.2 immediately implies that each probability distribution that agrees with the exchangeability OCM is exchangeable).

It is clear that the notion of a nonconformity measure in the exchangeability model is identical with that of a nonconformity measure as defined in Chap. 2, and so Proposition 2.4 (p. 27) is a special case of Theorem 8.1.

Generality and specificity

Let us say that a repetitive structure $M_2 = (\Sigma^2, \mathbf{Z}, (t_n^2), (P_n^2))$ is *more specific* than a repetitive structure $M_1 = (\Sigma^1, \mathbf{Z}, (t_n^1), (P_n^1))$ if there exists a sequence of measurable functions $f_n : \Sigma^1 \rightarrow \Sigma^2$ such that

- $t_n^2(z_1, \dots, z_n) = f_n(t_n^1(z_1, \dots, z_n))$ for all n and all data sequences $(z_1, \dots, z_n) \in \mathbf{Z}^n$;
- for each n and each $\sigma^2 \in t_n^2(\mathbf{Z}^n)$, the function $P_n^1(\cdot | \sigma^1)$, $\sigma^1 \in f_n^{-1}(\sigma^2)$, is a version of the conditional probability given that $t_n^1(z_1, \dots, z_n) = \sigma^1$ in the probability space $(\mathbf{Z}^n, P_n^2(\cdot | \sigma^2))$.

The first condition says that the statistics t_n^1 are more complete summaries of the data sequence z_1, \dots, z_n than the statistics t_n^2 are (since a summary preserves all useful information in the data sequence, this means that t_n^1 contains more noise, assuming both models are accepted). The second condition says that the probability distributions P_n^1 can be obtained from P_n^2 by conditioning on the more complete information.

The fact that M_2 is more specific than M_1 will be denoted $M_1 \preceq M_2$, and will also be expressed by saying that M_1 is *more general* than M_2 .

It is clear that if a probability distribution on \mathbf{Z}^∞ agrees with a repetitive structure, it will agree with a more general repetitive structure. As we know, a repetitive structure formalizes the assumption we are willing to make about Reality, and this assumption weakens as we replace a repetitive structure by a more general structure.

For simplicity, we did not state some results of Chap. 2 in their full generality. It is true that inductive and Mondrian conformal predictors are valid in the exchangeability model. But more than this is true: they are valid under weaker models, which will be considered in the following subsections.

Inductive-exchangeability models

As in §4.1, let m_1, m_2, \dots be a strictly increasing sequence, finite or infinite, of positive integers; if the sequence is finite, say m_1, \dots, m_r , we set $m_{r+1} := \infty$. For each such sequence m_1, m_2, \dots we can define the corresponding *inductive-exchangeability model* $(\Sigma, \mathbf{Z}, (t_n), (P_n))$, where:

- Σ is the set of finite sequences of bags of elements of \mathbf{Z} ;

- the summary $t_n(z_1, \dots, z_n)$ is defined as the sequence

$$\left(\{z_1, \dots, z_{m_1}\}, \{z_{m_1+1}, \dots, z_{m_2}\}, \dots, \right. \\ \left. \{z_{m_{k-1}+1}, \dots, z_{m_k}\}, \{z_{m_k+1}, \dots, z_n\} \right),$$

where k is such that $m_k < n \leq m_{k+1}$;

- for each $\sigma \in \Sigma_n$, $P_n(\cdot | \sigma)$ is defined as the uniform probability distribution on the set of $m_1! \cdots m_k!(n - m_k)!$ (with k defined as in the previous item) sequences obtained from σ by ordering its bags in different ways.

It is clear that inductive conformal predictors are conformal predictors in the inductive-exchangeability model (although not all conformal predictors in the inductive-exchangeability model are inductive conformal predictors); therefore, Proposition 4.1 is a special case of Theorem 8.1.

It is also clear that each inductive-exchangeability model is more general than the exchangeability model with the same example space. (The role of the hyphen is to emphasize that inductive-exchangeability models are not instances of exchangeability models, unless the sequence m_1, m_2, \dots is empty.) It is easy to see that an inductive-exchangeability model is strictly more general if m_1, m_2, \dots is not empty: if, e.g., m_1, m_2, \dots has only one term m , any product $Q_1^m \times Q_2^\infty$, where $Q_1, Q_2 \in \mathbf{P}(\mathbf{Z})$, will agree with the inductive-exchangeability model, whereas it will agree with the exchangeability model only if $Q_1 = Q_2$. The inductive-exchangeability model does not appear to be an interesting generalization of the exchangeability model (indeed, if $Q_1 \neq Q_2$, the inductive conformal predictor will be still valid, but its efficiency is likely to suffer), but in the next subsection we will see that Mondrian-exchangeability models can be quite useful.

Mondrian-exchangeability models

Following §4.5, fix a taxonomy $\kappa : \mathbb{N} \times \mathbf{Z} \rightarrow K$. The corresponding *Mondrian-exchangeability model* $(\Sigma, \mathbf{Z}, (t_n), (P_n))$ is defined as follows:

- Σ is the Cartesian product of K^* and the family of all mappings of the type $K \rightarrow \mathbf{Z}^{(*)}$:

$$\Sigma := K^* \times (\mathbf{Z}^{(*)})^K;$$

- the summary $t_n(z_1, \dots, z_n)$ is defined as the sequence

$$\left((\kappa(1, z_1), \dots, \kappa(n, z_n)), (k \in K \mapsto \{z_i : i \in \{1, \dots, n\}, \kappa(i, z_i) = k\}) \right);$$

- let $\sigma \in \Sigma_n$ consist of a sequence of categories k_1, \dots, k_n and a family of bags $(B_k : k \in K)$ of examples, such that each $k \in K$ occurs $|B_k|$ times in the sequence k_1, \dots, k_n ; $P_n(\cdot | \sigma)$ is then defined as the uniform probability distribution on the set of

$$\prod_{k \in K} |B_k|!$$

sequences z_1, \dots, z_n obtained from σ by ordering its bags in different ways and putting the elements of each ordered bag B_k consecutively in the places occupied by k in the sequence k_1, \dots, k_n .

Proposition 4.10 is a special case of Theorem 8.2: indeed, the latter shows that p_1, \dots, p_N are distributed as \mathbf{U}^N given the observed categories k_1, \dots, k_N .

It is easy to see that the notion of generality for repetitive structures (p. 197) as applied to Mondrian-exchangeability models agrees with the notion of generality for Mondrian taxonomies (p. 116).

Let us now discuss, for concreteness, what we called “label-conditional” MCPs in Chap. 4. To see that the Mondrian-exchangeability model is less restrictive in an important way than the exchangeability model, consider (following Ryabko 2003) the problem of hand-written character recognition. Suppose the stream of characters to be recognized comes from the user writing a letter. The exchangeability model is grossly wrong: the sequence of characters in a typical letter is far from exchangeable. If, however, we define the category of an example to be its label, the Mondrian-exchangeability model can be close to being correct: different instances of the character “a”, for example, can be almost exchangeable (even conditionally on the other characters and the way they are represented).

Mondrian-exchangeability models add another dimension to the usual question “what is exchangeable with what?” As in the case of Venn predictors, we want, on one hand, our predictions to be as specific as possible (which creates pressure on the categories to become smaller) and, on the other hand, we need enough statistics for each category (which resists the pressure). The pressure slightly increases, since smaller categories mean a weaker assumption about Reality.

8.5 Gaussian model

The inductive-exchangeability and Mondrian-exchangeability models are generalizations of the exchangeability model; we will now go in the other direction, considering a stronger model than exchangeability.

In the *Gaussian model*, $\mathbf{Z} := \mathbb{R}$, the statistics are

$$t_n(z_1, \dots, z_n) := (\bar{z}_n, \hat{\sigma}_n),$$

$$\bar{z}_n := \frac{1}{n} \sum_{i=1}^n z_i, \quad \hat{\sigma}_n^2 := \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z}_n)^2 \quad (8.6)$$

(except that $\hat{\sigma}_1 := 0$), and $P_n(dz_1, \dots, dz_n \mid \sigma_n)$ is the uniform distribution on $t_n^{-1}(\sigma_n)$ (in other words, it is the uniform distribution on the $(n-2)$ -dimensional sphere in \mathbb{R}^n with centre $(\bar{z}_n, \dots, \bar{z}_n) \in \mathbb{R}^n$ of radius $\sqrt{n-1}\hat{\sigma}_n$ lying inside the hyperplane $\frac{1}{n}(z_1 + \dots + z_n) = \bar{z}_n$).

It is clear that there are many possible representations of essentially the same model; for example, we obtain an equivalent model if we replace (8.6) by

$$t_n(z_1, \dots, z_n) := \left(\sum_{i=1}^n z_i, \sum_{i=1}^n z_i^2 \right). \quad (8.7)$$

Let us first find the forward functions in the on-line compression representation of the Gaussian model. It is easy to check that the updating formulae for \bar{z}_n and $\hat{\sigma}_n$ are

$$\begin{aligned} \bar{z}_n &= \frac{n-1}{n} \bar{z}_{n-1} + \frac{1}{n} z_n, \\ \hat{\sigma}_n^2 &= \frac{n-2}{n-1} \hat{\sigma}_{n-1}^2 + \frac{1}{n} (z_n - \bar{z}_{n-1})^2; \end{aligned}$$

this defines the forward functions F_n . The expression for the backward kernels is much more complicated, and we do not give it explicitly; it can be derived from the fact that (8.9) below has Student's t -distribution.

Let us now explicitly find the prediction set for the Gaussian model and nonconformity measure

$$A_n(\sigma_{n-1}, z_n) = A_n((\bar{z}_{n-1}, \hat{\sigma}_{n-1}), z_n) := |z_n - \bar{z}_{n-1}| \quad (8.8)$$

(it is easy to check that this nonconformity measure is equivalent, in the sense of leading to the same p-values, to $|z_n - \bar{z}_n|$, as well as to several other natural expressions, including (8.9)). Under $P_n(dz_1, \dots, dz_n | (\bar{z}_n, \hat{\sigma}_n))$ and assuming $n > 2$, the expression

$$\sqrt{\frac{n-1}{n}} \frac{|z_n - \bar{z}_{n-1}|}{\hat{\sigma}_{n-1}} \quad (8.9)$$

has the t -distribution with $n-2$ degrees of freedom. (This fact is proved in, e.g., Cramér 1946, §29.4, where it is assumed, however, that z_1, \dots, z_n are independent and have the same normal distribution. The latter assumption may be replaced by our assumption of the uniform distribution; for a general argument, see the proof of Proposition 8.4 below.) Let $t_{\delta,k}$ be the value defined by $\mathbb{P}\{\xi \geq t_{\delta,k}\} = \delta$ with ξ having the t -distribution with k degrees of freedom. We can see that the prediction set Γ_n^ϵ corresponding to nonconformity measure (8.8) is the interval consisting of z such that

$$|z - \bar{z}_{n-1}| \leq t_{\epsilon/2, n-2} \sqrt{\frac{n}{n-1}} \hat{\sigma}_{n-1}. \quad (8.10)$$

We obtained the usual prediction set based on the t -test (as in Baker 1935, Wilks 1941, and, implicitly, Fisher 1925); now, however, we can see that the errors of this standard procedure (applied in the on-line fashion) are independent.

Some of the facts mentioned in this subsection will be proved in the following one.

Gauss linear model

We will now consider a rich extension of the Gaussian model. In the repetitive-structure representation $(\Sigma, \mathbf{Z}, (t_n), (P_n))$ of the *Gauss linear model* the example space is of the regression type, $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$, with the label space being the real line $\mathbf{Y} := \mathbb{R}$ and the object space being the p -dimensional Euclidean space, $\mathbf{X} := \mathbb{R}^p$. The statistics are

$$t_n(x_1, y_1, \dots, x_n, y_n) := \left(x_1, \dots, x_n, \sum_{i=1}^n y_i x_i, \sum_{i=1}^n y_i^2 \right) \quad (8.11)$$

(so Σ can be set to $\mathbf{X}^* \times \mathbb{R}^p \times \mathbb{R}$), and $P_n(\cdot | \sigma_n)$ is the uniform probability distribution on the sphere $t_n^{-1}(\sigma_n)$ (we consider a point to be a sphere; typically $t_n^{-1}(\sigma_n)$ will be a point unless $n > p$).

The Gaussian model in the form (8.7) is a special case (using a different notation, z_i for y_i) corresponding to $p = 1$ and x_i restricted to $x_i = 1$, $i = 1, 2, \dots$. Using $\sum_{i=1}^n y_i x_i$ rather than $\sum_{i=1}^n y_i$ reflects the possibility that y_i can depend on x_i .

The probability distribution of z_1, z_2, \dots under the linear regression statistical model

$$y_n = w \cdot x_n + \xi_n, \quad (8.12)$$

where $w \in \mathbb{R}^p$ is a constant vector and ξ_n are independent random variables with the same zero-mean normal distribution, always agrees with the Gauss linear model. (The model (8.12) was already considered in Chap. 2: cf. (2.37) on p. 35.)

Our next proposition and its proof will use the following notation: \hat{y}_i^n is the least squares prediction for the object x_i based on the examples z_1, \dots, z_n ; \hat{y}_n is a shorthand for \hat{y}_n^{n-1} ; X_l , $l = 1, 2, \dots$, is the $l \times p$ matrix whose i th row is x'_i , $i = 1, \dots, l$; and

$$\hat{\sigma}_l^2 := \frac{1}{l-p} \sum_{i=1}^l (y_i - \hat{y}_i^l)^2$$

is the standard estimate of the variance of the Gaussian noise ξ_n in (8.12) from the first l examples.

Proposition 8.4. *The conformal predictor determined by the nonconformity measure*

$$A(\sigma, (x, y)) := |y - \hat{y}|,$$

where \hat{y} is the least squares prediction of the x 's label y based on the examples summarized by σ , is given, for $n > p+1$ satisfying $\text{rank}(X_{n-1}) = p$, by the formula

$$\Gamma_n^\epsilon = [\hat{y}_n - \mathbf{t}_{\epsilon/2, n-p-1} V_n, \hat{y}_n + \mathbf{t}_{\epsilon/2, n-p-1} V_n], \quad (8.13)$$

where

$$V_n := \sqrt{1 + x'_n (X'_{n-1} X_{n-1})^{-1} x_n \hat{\sigma}_{n-1}}.$$

Confidence predictor (8.13), generalizing (8.10), will be called the *Student predictor*.

Proof. It is a standard fact (see, e.g., Stuart et al. 1999, §32.10) that $(y_n - \hat{y}_n)/V_n$ has the t -distribution with $n - p - 1$ degrees of freedom; this assumes, however, the standard model (8.12) rather than the uniform conditional distribution of the Gauss linear model. Let us check that $(y_n - \hat{y}_n)/V_n$ will still have the t -distribution with $n - p - 1$ degrees of freedom under the uniform conditional distribution.

First note that $(y_n - \hat{y}_n)/V_n$ can be rewritten so that it depends on y_1, \dots, y_n only through the n -residuals $y_i - \hat{y}_i^n$ (i.e., residuals computed from all n examples z_1, \dots, z_n). Indeed, a standard statistical result (Montgomery et al. 2001, (4.12)) shows that

$$\hat{\sigma}_{n-1}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i^n)^2 - (y_n - \hat{y}_n^n)^2 / (1 - x_n'(X_n' X_n)^{-1} x_n)}{n - p - 1}; \quad (8.14)$$

another standard result (Montgomery et al. 2001, (4.11); already used on p. 34) shows that

$$y_n - \hat{y}_n = \frac{y_n - \hat{y}_n^n}{1 - x_n'(X_n' X_n)^{-1} x_n}. \quad (8.15)$$

Remember that $Y_n := (y_1, \dots, y_n)'$ is the vector of the first n labels and let $\hat{Y}_n := (\hat{y}_1^n, \dots, \hat{y}_n^n)'$ be the vector of the first n fitted values. According to the geometric interpretation of the least squares method in the standard model (8.12) (see, e.g., Draper and Smith 1998, Chaps. 20–21), the vector of n -residuals is distributed symmetrically around \hat{Y}_n in the space orthogonal to the estimation space $\{X_n w : w \in \mathbb{R}^p\}$. On the other hand, according to (8.11) and the definition of P_n , $P_n(\cdot | \sigma_n)$ is the uniform distribution on the sphere, of radius equal to the length of the vector of n -residuals, in the hyperplane orthogonal to the estimation space and passing through the projection \hat{Y}_n of Y_n onto the estimation space. Since the ratio $(y_n - \hat{y}_n)/V_n$ (expressed through the n -residuals $y_i - \hat{y}_i^n$) does not change if all n -residuals are multiplied by the same positive constant (and, therefore, its distribution does not change if the random vector of n -residuals is scaled to have a given length), we may replace the normal distribution of (8.12) by our uniform distribution $P_n(\cdot | \sigma_n)$.

The proof will be complete if we show that

$$\left| \frac{y_n - \hat{y}_n}{V_n} \right| = \frac{|y_n - \hat{y}_n|}{V_n}$$

is a bona fide nonconformity measure which monotonically increases as $|y_n - \hat{y}_n|$ increases for any fixed $\sigma_n := t_n(z_1, \dots, z_n)$. To see that $|y_n - \hat{y}_n|/V_n$ can be expressed through $\sigma_{n-1} := t_{n-1}(x_1, y_1, \dots, x_{n-1}, y_{n-1})$ (see (8.11)) and $z_n = (x_n, y_n)$, it suffices to remember that

$$\hat{y}_n = Y_{n-1}' X_{n-1} (X_{n-1}' X_{n-1})^{-1} x_n$$

(see (2.29) on p. 30) and notice that, in accordance with (8.14) and (8.15), V_n can also be expressed through σ_{n-1} and z_n :

$$\sum_{i=1}^n (y_i - \hat{y}_i^n)^2 = \|Y_n - \hat{Y}_n\|^2 = \|Y_n\|^2 - \|\hat{Y}_n\|^2 = \|Y_n\|^2 - \|H_n Y_n\|^2,$$

where $H_n = X_n(X'_n X_n)^{-1} X'_n$ is the hat matrix (2.31) (p. 31). Finally, we deduce from (8.14) and (8.15):

$$\begin{aligned} \frac{|y_n - \hat{y}_n|}{V_n} &\uparrow\uparrow \frac{|y_n - \hat{y}_n|}{\sqrt{C - c(y_n - \hat{y}_n)^2}} \uparrow\uparrow \frac{(y_n - \hat{y}_n)^2}{C - c(y_n - \hat{y}_n)^2} \\ \uparrow\downarrow \frac{C - c(y_n - \hat{y}_n)^2}{(y_n - \hat{y}_n)^2} &\uparrow\uparrow \frac{1}{(y_n - \hat{y}_n)^2} \uparrow\downarrow |y_n - \hat{y}_n| \uparrow\uparrow |y_n - \hat{y}_n|, \end{aligned}$$

where $C > 0$ and c are constants (for a fixed σ_n), $\uparrow\uparrow$ means “changes in the same direction as”, and $\uparrow\downarrow$ means “changes in the opposite direction to”. \square

The prediction interval (8.13) is standard (see, e.g., Montgomery et al. 2001, (3.54)), but our results add the usual extra feature: the independence of errors in the on-line setting.

Remark The methods of this subsection are applicable to time series, although only to the simplest ones: e.g., if

$$y_n = f(n) + \cos \frac{n-a}{T} + \xi_n$$

where $f(n)$ is a polynomial of a known degree p , T is a known constant (the period of the seasonal component), and ξ_n are independent and identically distributed zero-mean normal random variables, we can set

$$x_n := \left(1, n, \dots, n^p, \cos \frac{n}{T}, \sin \frac{n}{T}\right)$$

and use formula (8.13). Constructing conformal predictors in more interesting cases would require new methods.

Student predictor vs. ridge regression confidence machine

In this subsection we will investigate empirically the efficiency of RRCM. The idea is to run both the Student predictor and RRCM on a data set generated from an exchangeable probability distribution in the linear regression model (8.12). The RRCM “does not know” that the labels are generated from the simple parametric model (8.12), and so one would expect the RRCM to work worse than the Student predictor, which is tuned to (8.12)¹. An interesting

¹Of course, the situation is somewhat symmetrical in that the Student predictor “does not know” the data is exchangeable, but (8.12) appears more useful in predicting the labels than the exchangeability.

question is “how much worse?” We will see that the difference is surprisingly small for the Boston Housing data set, suitably “shuffled” (cf. §B.4) to conform to the Gauss linear and exchangeability assumptions.

First we discuss the details of “shuffling”. Suppose we have N examples (x_n, y_n) , $n = 1, \dots, N$, with $x_n \in \mathbb{R}^p$ and $y_n \in \mathbb{R}$, and we would like to generate another sequence of N examples $(x_n^*, y_n^*) = (x_n, y_n)$ from the uniform distribution $P_N(\cdot | \sigma_N)$, where $\sigma_N = t_N(x_1, y_1, \dots, x_N, y_N)$ is defined by (8.11). The procedure is:

- Let X be the $N \times p$ matrix with rows x'_n , $n = 1, \dots, N$, and Y be the column-vector $(y_1, \dots, y_N)'$ of length N . (We assume $\text{rank } X = p$.)
- Compute $Y_1 := HY$, where $H := X(X'X)^{-1}X'$ is the hat matrix.
- Compute the linear combination Y_2 with independent $\mathbf{N}_{0,1}$ coefficients of the vectors in an orthonormal basis of the linear space $\{y \in \mathbb{R}^N : X'y = 0\}$. (We will assume that $\|Y_2\| \neq 0$; this will be the case with probability one when $N > p$.)
- Set

$$Y^* := Y_1 + \frac{Y_2}{\|Y_2\|} \sqrt{\|Y\|^2 - \|Y_1\|^2}$$

and output the n th component of Y^* as y_n^* , $n = 1, \dots, N$.

Since Y_1 is the projection of Y onto the estimation space (the subspace of \mathbb{R}^N generated by the columns of X) and $Y_2/\|Y_2\|$ is a random vector of norm 1 in the orthogonal complement in \mathbb{R}^N of the estimation space, the validity of our procedure follows from $\|Y^*\| = \|Y\|$; the latter is obvious since Y_1 and Y_2 are orthogonal.

We first randomly permuted the elements of the Boston Housing data set (this step can be called “exchangeability shuffling”) and then applied the above procedure to the resulting sequence (“Gauss linear shuffling”). We then run the RRCM² and the Student predictor in the on-line fashion on this doubly shuffled data set; the results are shown in Figures 8.3 and 8.4, in the same format as in Chap. 2. On both graphs, the solid line shows, for each $n = 1, \dots, 506$, the median $M_n^{99\%}$ of the widths of the convex hulls $\text{co } \Gamma_i^{1\%}$ of the prediction sets $\Gamma_i^{1\%}$, $i = 1, \dots, n$; similarly, the dashed line shows $M_n^{95\%}$ and the dash-dot line shows $M_n^{80\%}$. (In the case of the Student predictor, Γ_i^ϵ are convex, and so $\text{co } \Gamma_i^\epsilon = \Gamma_i^\epsilon$.) The cumulative error lines look as usual (Figs. 8.5 and 8.6). The RRCM does not look much worse (and, surprisingly, one part of the graph for RRCM looks even better – the one corresponding to small n and confidence level 80%).

²The ridge coefficient for the RRCM was $a = 1$ and each attribute was linearly scaled to span the interval $[-1, 1]$ (or $[0, 0]$, if its maximum and minimum values coincided).

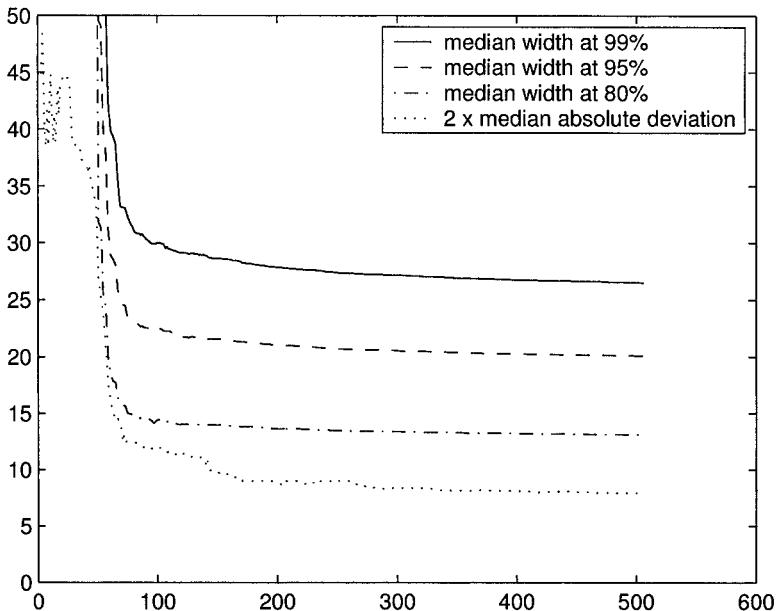


Fig. 8.3. The on-line performance of the Student predictor on the doubly shuffled Boston Housing data set

8.6 Markov model

We have already considered two classes of on-line compression models that go beyond exchangeability in interesting ways: the Mondrian-exchangeability models of §8.4 and the Gauss linear models of §8.5. In this section we will consider a third class of non-exchangeable models (translation of the probabilistic notion of Markov chain into our framework).

We are not so much interested in prediction for Markov chains *per se*: the corresponding statistical model is a regular finite-dimensional model (in the case of binary Markov chains, which will be our focus of attention, there are only two parameters to be learned: the probability of 1 after 0 and the probability of 1 after 1), and valid and efficient inductive prediction in this situation is easy. (Details will be given in §10.1.) But this simple example will demonstrate an important limitation of transductive prediction (at least, when it is used in the most direct way), which can well show up in more interesting applications.

The unusual feature of transductive prediction when applied to Markov chains is an apparent efficiency/validity trade-off. The method of this section is different from the one used in §7.3, where we also treated Markov chains. It is interesting that the two methods give different results. The method of this section is valid in a stronger sense (the whole sequence of errors is Bernoulli),

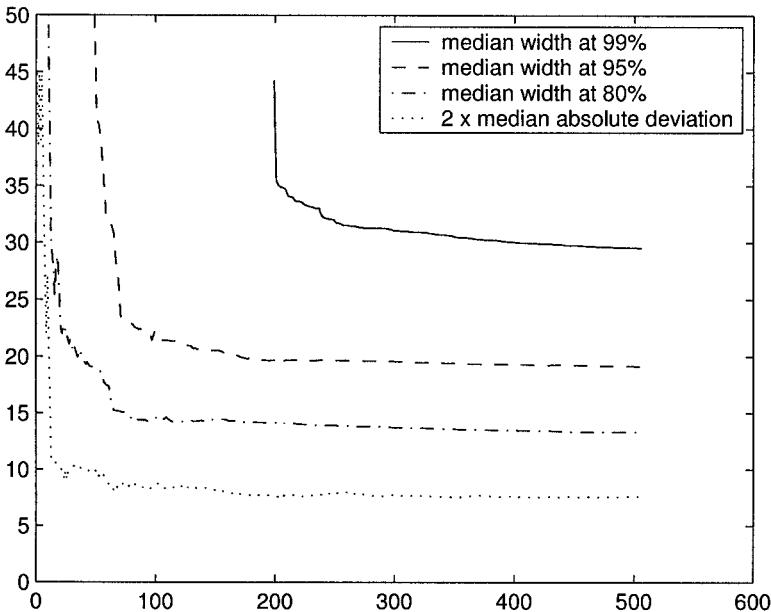


Fig. 8.4. The on-line performance of RRCM on the doubly shuffled Boston Housing data set

whereas the method of §7.3 appears to be more efficient. This will be discussed in detail at the end of §8.8.

In this section we always assume that the example space \mathbf{Z} is finite (often binary, $\mathbf{Z} = \{0, 1\}$). We start by giving some basic definitions of graph theory in a convenient for us form.

There are two natural variants of the notion of a directed graph; we will use “digraph” and “semi-Markov graph” as technical terms. A *digraph* with a vertex set V and an arc set E is given by two mappings: the *tail mapping* $\text{tail} : E \rightarrow V$ and the *head mapping* $\text{head} : E \rightarrow V$. The digraph is drawn by representing each vertex by a dot and representing each arc $e \in E$ by an arrow leading from $\text{tail}(e)$ to $\text{head}(e)$. A *semi-Markov graph* with a vertex set V is specified by a bag of elements of V^2 ; each element of the bag is called an *arc*. Again, a semi-Markov graph is drawn by representing each vertex by a dot and representing each arc (v_1, v_2) by an arrow from v_1 to v_2 .

An *Eulerian path* in a digraph is a sequence of alternating vertices and edges $v_1, e_1, v_2, e_2, \dots, v_n, e_n$ such that each e_i , $i = 1, \dots, n$, leads from v_i to v_{i+1} (v_{n+1} is understood to be v_1) and the arcs e_1, \dots, e_n form an ordering of E (without repetitions). This notion is standard; in this section, however, the following will be more useful. An *Eulerian path* in a semi-Markov graph is a sequence of vertices v_1, v_2, \dots, v_n such that the bag

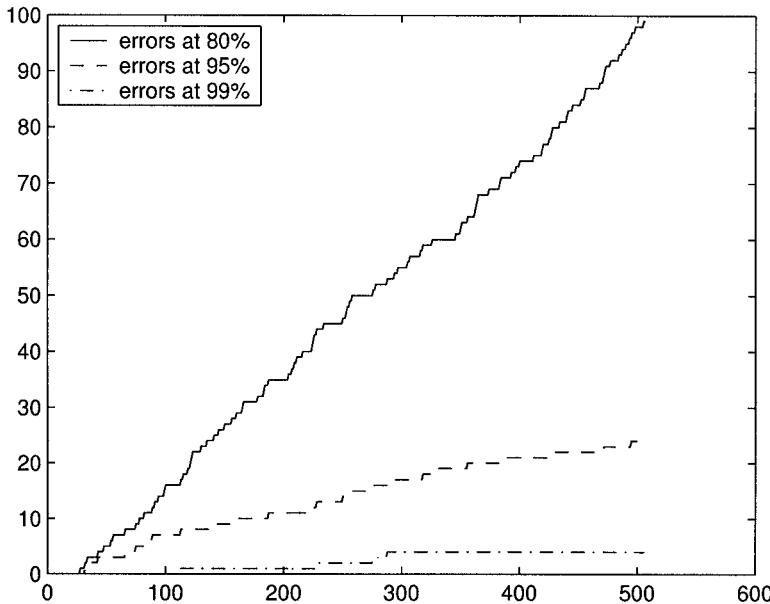


Fig. 8.5. The cumulative numbers of errors at the given confidence levels for the Student predictor run on-line on the doubly shuffled Boston Housing data set

$\{(v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, v_n)\}$ coincides with the bag that specifies the semi-Markov graph; we will say that this path is *from* v_1 *to* v_n .

Intuitively, the difference between digraphs and Markov semi-graphs is that for the latter the arcs with the same tail and head are indistinguishable (for example, we do not distinguish two Eulerian paths that only differ in the order in which two such arcs are passed). The *underlying digraph* of a Markov semi-graph will have the same structure but all its arcs will be considered to have their own identity.

The following notation for digraphs and Markov semi-graphs will be used: $\text{in}(v)/\text{out}(v)$ stand for the number of arcs entering/leaving vertex v ; $n_{u,v}$ is the number of arcs leading from vertex u to vertex v .

The *Markov summary* of a data sequence $z_1 \dots z_n$ is the following Markov semi-graph with two vertices marked:

- the set of vertices is \mathbf{Z} (the state space of the Markov chain);
- the vertex z_1 is marked as the *source* and the vertex z_n is marked as the *sink* (these two vertices are not necessarily distinct);
- the arcs of the Markov semi-graph are the transitions $z_i z_{i+1}$, $i = 1, \dots, n-1$; the arc $z_i z_{i+1}$ has z_i as its tail and z_{i+1} as its head.

It is clear that in any Markov summary all vertices v satisfy $\text{in}(v) = \text{out}(v)$ with the possible exception of the source and sink (unless they coincide), for which we then have $\text{out}(\text{source}) = \text{in}(\text{source}) + 1$ and $\text{in}(\text{sink}) = \text{out}(\text{sink}) + 1$.

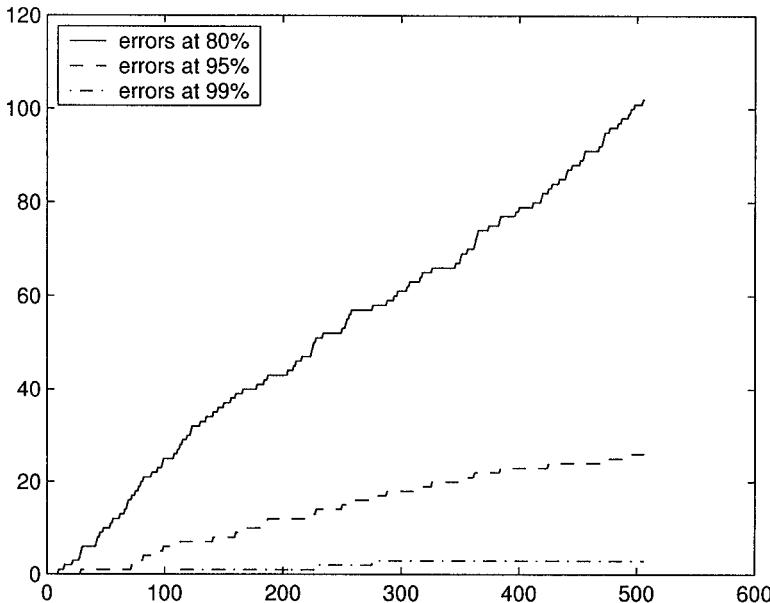


Fig. 8.6. The cumulative numbers of errors at the given confidence levels for RRCM run on-line on the doubly shuffled Boston Housing data set

We will call a Markov semi-graph with two vertices marked as the source and sink a *Markov graph* if it satisfies this property.

The repetitive-structure representation $(\Sigma, \mathbf{Z}, (t_n), (P_n))$ of the Markov model is:

- \mathbf{Z} is a finite set; its elements (examples) are also called *states*;
- Σ is the set of all Markov graphs with the vertex set \mathbf{Z} ;
- $t_n(z_1, \dots, z_n)$ is the Markov summary of the data sequence $z_1 \dots z_n$;
- for each $\sigma \in \Sigma_n$, $P_n(\sigma)$ is the uniform probability distribution on the set of Eulerian paths from the source to the sink in σ .

The Markov model can also be defined directly as an on-line compression model $(\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$:

- \mathbf{Z} is a finite set;
- Σ is the set of all Markov graphs with the vertex set \mathbf{Z} extended by adding a new element \square (say, the empty set);
- $F_1(\square, z)$ is the Markov graph with no arcs and with both source and sink at z ; $F_n(\sigma, z)$, $n > 1$, is the Markov graph obtained from σ by adding an arc from σ 's sink to z and making z the new sink;
- let $\sigma \downarrow z$, where σ is a Markov graph and z is one of σ 's vertices, be the Markov graph obtained from σ by removing an arc from z to σ 's sink ($\sigma \downarrow z$ does not exist if there is no arc from z to σ 's sink) and moving the sink

to z , and let $\#\sigma$ be the number of Eulerian paths from the source to the sink in a Markov graph σ ; $B_n(\sigma)$ generates $(\sigma \downarrow z, \text{sink})$ with probability $\#(\sigma \downarrow z)/\#\sigma$, where sink is σ 's sink and z ranges over the states for which $\sigma \downarrow z$ is defined.

Notice that any Markov probability distribution on \mathbf{Z}^∞ agrees with the Markov model.

We will take

$$A_n(\sigma, z) := B_n(\{(\sigma, z)\} \mid F_n(\sigma, z)) \quad (8.16)$$

as the conformity measure (intuitively, lower probability makes an example less conforming). To give a computationally efficient representation of the conformal transducer corresponding to this conformity measure, we need the following two graph-theoretic results, versions of the BEST theorem and the Matrix-Tree theorem, respectively.

Lemma 8.5. *In any Markov graph σ with the set of vertices V the number of Eulerian paths from the source to the sink equals*

$$T(\sigma) \frac{\text{out}(\text{sink}) \prod_{v \in V} (\text{out}(v) - 1)!}{\prod_{u,v \in V} n_{u,v}!},$$

where $T(\sigma)$ is the number of spanning out-trees in the underlying digraph rooted at the source.

Lemma 8.6. *To find the number $T(\sigma)$ of spanning out-trees rooted at the source in the underlying digraph of a Markov graph σ with vertices z_1, \dots, z_n (z_1 being the source),*

- create the $n \times n$ matrix with the elements $a_{i,j} = -n_{z_i, z_j}$;
- change the diagonal elements so that each column sums to 0;
- compute the co-factor of $a_{1,1}$.

These two lemmas immediately follow from Theorems VI.24 and VI.28 in Tutte 2001. To derive Lemma 8.5, notice that counting Eulerian paths from the source to the sink in a Markov graph reduces to counting Eulerian paths in the underlying digraph with an added arc from the sink to the source.

It is now easy to obtain an explicit formula for prediction in the binary case $\mathbf{Z} = \{0, 1\}$. First we notice that, for $n > 1$,

$$B_n(\{(\sigma \downarrow z, \text{sink})\} \mid \sigma) = \frac{\#(\sigma \downarrow z)}{\#\sigma} = \frac{T(\sigma \downarrow z) n_{z, \text{sink}}}{T(\sigma) \text{out}(\text{sink})}$$

(all $n_{u,v}$ refer to the numbers of arcs in σ and sink is σ 's sink; we set $\#(\sigma \downarrow z) = T(\sigma \downarrow z) := 0$ when $\sigma \downarrow z$ does not exist). The following simple corollary from the last formula is sufficient for computing the probabilities B_n in the binary case:

$$B_n(\{(\sigma \downarrow \text{sink}, \text{sink})\} \mid \sigma) = \frac{n_{\text{sink}, \text{sink}}}{\text{out}(\text{sink})}.$$

This gives us the following formulas for the conformal predictor in the binary Markov model (remember that the conformity measure is (8.16)). Suppose the current summary is given by a Markov graph with $n_{i,j}$ arcs going from vertex i to vertex j ($i, j \in \{0, 1\}$) and let $f : [0, 1] \rightarrow [0, 1]$ be the function that squashes $[0.5, 1]$ to 1:

$$f(p) := \begin{cases} p & \text{if } p < 0.5 \\ 1 & \text{otherwise.} \end{cases}$$

If the current sink is 0, the p-value corresponding to the next example 0 is

$$f\left(\frac{n_{0,0} + 1}{n_{0,0} + n_{0,1} + 1}\right)$$

and the p-value corresponding to the next example 1 is (with $0/0 := 1$)

$$f\left(\frac{n_{1,0}}{n_{1,0} + n_{1,1}}\right). \quad (8.17)$$

If the current sink is 1, the p-value corresponding to the next example 1 is

$$f\left(\frac{n_{1,1} + 1}{n_{1,1} + n_{1,0} + 1}\right)$$

and the p-value corresponding to the next example 0 is (with $0/0 := 1$)

$$f\left(\frac{n_{0,1}}{n_{0,1} + n_{0,0}}\right). \quad (8.18)$$

Figure 8.7 shows the result of a computer simulation; as expected, the error line is close to the straight line with the slope close to the significance level.

In conclusion, we notice the profound difference between the formulas of this section and the recipe of §7.3. According to the latter, for example, we should have

$$f\left(\frac{n_{0,1} + 1}{n_{0,0} + n_{0,1} + 1}\right) \quad (8.19)$$

instead of (8.17); expression (8.19) is closer to (8.18). This will be discussed, at an informal level, later in §8.8 (from p. 220).

The approach of this section works well in the case where the Markov chain is expected to be symmetric or close to symmetric but we want a guarantee that validity will not be lost even if the Markov chain is very far from symmetry. (If we are certain that the Markov chain is symmetric, it is best to use the OCM with the statistics

$$t_n(z_1, \dots, z_n) := (|\{i = 1, \dots, n-1 : z_i \neq z_{i+1}\}|, z_n),$$

as in Lauritzen 1988, p. 45, but this model easily reduces to the binary exchangeability model for $z'_i := \mathbb{I}_{z_i \neq z_{i+1}}$.)

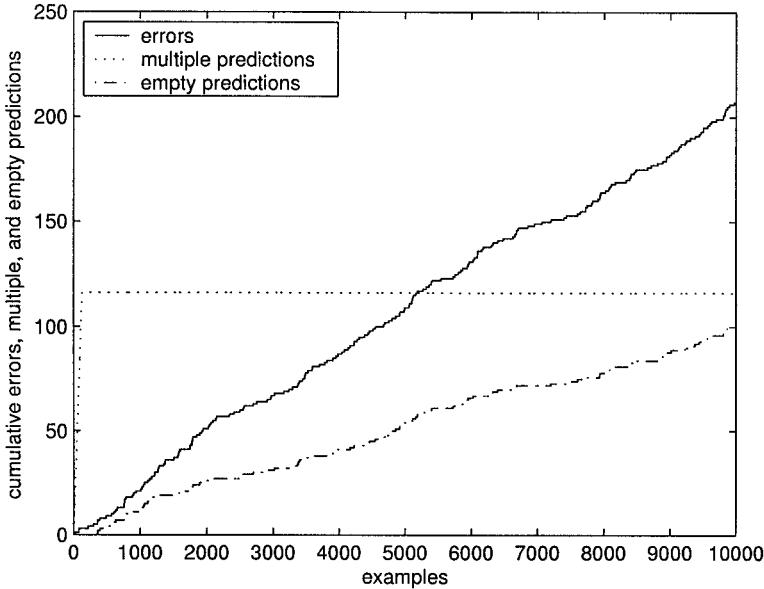


Fig. 8.7. Conformal predictor predicting the binary Markov chain with the following transition probabilities: 0 is followed by 1 with probability 1% and 1 is followed by 0 with the same probability 1%. The significance level is 2%; the cumulative numbers of errors, multiple, and empty prediction sets are shown

8.7 Proof of Theorem 8.2

First we explain the basic idea of the proof. To show that (p_1, \dots, p_N) is distributed as \mathbf{U}^N , we use the standard idea of reversing the time (see, e.g., the proof of de Finetti's theorem in Schervish 1995). Let P be the distribution on \mathbf{Z}^N generating the examples; it is assumed to agree with the OCM. We can imagine that the data sequence (z_1, \dots, z_N) is generated in two steps: first, the summary σ_N is generated from some probability distribution (namely, the image of the distribution P under the mapping t_N), and then the data sequence (z_1, \dots, z_N) is chosen randomly from $P_N(\cdot | \sigma_N)$. Already the second step ensures that, conditionally on knowing σ_N (and, therefore, unconditionally), the sequence (p_N, \dots, p_1) is distributed as \mathbf{U}^N . Indeed, roughly speaking (i.e., ignoring borderline effects), p_N will be the p-value corresponding to the statistic A_N and so distributed, at least approximately, as \mathbf{U} (see, e.g., Cox and Hinkley 1974, §3.2); when the pair (σ_{N-1}, z_N) is disclosed, the value p_N will be settled; conditionally on knowing σ_{N-1} and z_N , p_{N-1} will also be distributed as \mathbf{U} , and so on.

We start the formal proof by defining the σ -algebra \mathcal{G}_n , $n = 0, 1, \dots, N$, as the one on the sample space $(\mathbf{Z} \times [0, 1])^N$ generated by the random elements $\sigma_n := t_n(z_1, \dots, z_n), z_{n+1}, \tau_{n+1}, z_{n+2}, \tau_{n+2}, \dots, z_N, \tau_N$. In particular, \mathcal{G}_0 (the

most informative σ -algebra) coincides with the original σ -algebra on $(\mathbf{Z} \times [0, 1])^N; \mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \dots \supseteq \mathcal{G}_N$.

Fix a smoothed conformal transducer f ; it will usually be left implicit in our notation. Let p_n be the random variable $f(z_1, \tau_1, \dots, z_n, \tau_n)$ for each $n = 1, \dots, N$; \mathbb{P} will refer to the probability distribution $P \times \mathbf{U}^N$ (over examples z_n and random numbers τ_n) and \mathbb{E} to the expectation w.r. to \mathbb{P} . It will be convenient to write $\mathbb{P}_{\mathcal{G}}(A)$ and $\mathbb{E}_{\mathcal{G}}(\xi)$ for the conditional probability $\mathbb{P}(A | \mathcal{G})$ and expectation $\mathbb{E}(\xi | \mathcal{G})$, respectively. The proof will be based on the following lemma.

Lemma 8.7. *For any trial $n = 1, \dots, N$ and any $\epsilon \in [0, 1]$,*

$$\mathbb{P}_{\mathcal{G}_n} \{p_n \leq \epsilon\} = \epsilon. \quad (8.20)$$

Proof. Let us fix a summary σ_n of the first n examples $(z_1, \dots, z_n) \in \mathbf{Z}^n$; we will omit the condition “ $| \sigma_n$ ”. For every pair $(\tilde{\sigma}, \tilde{z})$ from $F_n^{-1}(\sigma_n)$ define

$$\begin{aligned} p^+(\tilde{\sigma}, \tilde{z}) &:= B_n \{(\sigma, z) : A_n(\sigma, z) \geq A_n(\tilde{\sigma}, \tilde{z})\}, \\ p^-(\tilde{\sigma}, \tilde{z}) &:= B_n \{(\sigma, z) : A_n(\sigma, z) > A_n(\tilde{\sigma}, \tilde{z})\}. \end{aligned}$$

It is clear that always $p^- \leq p^+$.

Let us say that a pair $(\tilde{\sigma}, \tilde{z})$ is

- *strange* if $p^+(\tilde{\sigma}, \tilde{z}) \leq \epsilon$
- *conforming* if $p^-(\tilde{\sigma}, \tilde{z}) > \epsilon$
- *borderline* if $p^-(\tilde{\sigma}, \tilde{z}) \leq \epsilon < p^+(\tilde{\sigma}, \tilde{z})$.

We will use the notation $p^- := p^-(\tilde{\sigma}, \tilde{z})$ and $p^+ := p^+(\tilde{\sigma}, \tilde{z})$ where $(\tilde{\sigma}, \tilde{z})$ is any borderline example (if there are no borderline examples, $p^- = p^+$ can be defined as $\inf p^-(\tilde{\sigma}, \tilde{z})$ over the conforming $(\tilde{\sigma}, \tilde{z})$ or as $\sup p^+(\tilde{\sigma}, \tilde{z})$ over the strange $(\tilde{\sigma}, \tilde{z})$). Notice that the B_n -measure of strange examples is p^- , the B_n -measure of conforming examples is $1 - p^+$, and the B_n -measure of borderline examples is $p^+ - p^-$.

By the definition of a smoothed conformal transducer, $p_n \leq \epsilon$ if the pair (σ_{n-1}, z_n) is strange, $p_n > \epsilon$ if the pair is conforming, and $p_n \leq \epsilon$ with probability

$$\frac{\epsilon - p^-}{p^+ - p^-}$$

(with, say, $0/0 := 0$) if the pair is borderline; indeed, in the latter case, as

$$p_n = p^- + \tau_n(p^+ - p^-),$$

$p_n \leq \epsilon$ is equivalent to

$$\tau_n \leq \frac{\epsilon - p^-}{p^+ - p^-}.$$

Therefore, the overall probability that $p_n \leq \epsilon$ is

$$p^- + (p^+ - p^-) \frac{\epsilon - p^-}{p^+ - p^-} = \epsilon. \quad \square$$

The other basic result that we will need is the following lemma.

Lemma 8.8. *For any trial $n = 1, \dots, N$, p_n is \mathcal{G}_{n-1} -measurable.*

Proof. This follows from the definition, (8.4) on p. 193: p_n is defined in terms of σ_{n-1} , z_n and τ_n . The only technicality that might not be immediately obvious is that the function

$$B_n(\{A_n > c\} | \sigma)$$

of $c \in \mathbb{R}$ and $\sigma \in \Sigma$ is measurable. Let $C \in \mathbb{R}$. The set

$$\{(c, \sigma) : B_n(\{A_n > c\} | \sigma) > C\} \quad (8.21)$$

is measurable since it can be represented as

$$\bigcup_{d \in \mathbb{Q}} (0, d) \times \Sigma_d,$$

where \mathbb{Q} is the set of rational numbers and Σ_c is the set of σ satisfying the outer inequality in (8.21). \square

First we prove that, for any $n = 1, \dots, N$ and any $\epsilon_1, \dots, \epsilon_n \in [0, 1]$,

$$\mathbb{P}_{\mathcal{G}_n}\{p_n \leq \epsilon_n, \dots, p_1 \leq \epsilon_1\} = \epsilon_n \cdots \epsilon_1 \quad \text{a.s.} \quad (8.22)$$

The proof is by induction on n . For $n = 1$, (8.22) is a special case of Lemma 8.7. For $n > 1$ we obtain, making use of Lemmas 8.7 and 8.8, properties 1 and 2 of conditional expectations (see p. 279), and the inductive assumption:

$$\begin{aligned} \mathbb{P}_{\mathcal{G}_n}\{p_n \leq \epsilon_n, \dots, p_1 \leq \epsilon_1\} \\ = \mathbb{E}_{\mathcal{G}_n}(\mathbb{E}_{\mathcal{G}_{n-1}}(\mathbb{I}_{p_n \leq \epsilon_n} \mathbb{I}_{p_{n-1} \leq \epsilon_{n-1}, \dots, p_1 \leq \epsilon_1})) \\ = \mathbb{E}_{\mathcal{G}_n}(\mathbb{I}_{p_n \leq \epsilon_n} \mathbb{E}_{\mathcal{G}_{n-1}}(\mathbb{I}_{p_{n-1} \leq \epsilon_{n-1}, \dots, p_1 \leq \epsilon_1})) \\ = \mathbb{E}_{\mathcal{G}_n}(\mathbb{I}_{p_n \leq \epsilon_n} \epsilon_{n-1} \cdots \epsilon_1) = \epsilon_n \epsilon_{n-1} \cdots \epsilon_1 \end{aligned}$$

almost surely.

By property 2, (8.22) immediately implies

$$\mathbb{P}\{p_N \leq \epsilon_N, \dots, p_1 \leq \epsilon_1\} = \epsilon_N \cdots \epsilon_1.$$

Remark In our definitions of conformal transducer, conformal predictor, etc., we assumed that the same random number τ_n is used for every potential label y of the new object x_n . In fact, assuming \mathbf{Y} is finite, we can also use a separate random number τ_n^y for each $y \in \mathbf{Y}$, with the random numbers τ_n^y , $n = 1, 2, \dots$, $y \in \mathbf{Y}$, independent. On the other hand, an arbitrary correlation between τ_n^y , $y \in \mathbf{Y}$, can be allowed; Theorems 8.1 and 8.2 will continue to hold as long as the random numbers $\tau_n^{y_n}$, $n = 1, 2, \dots$, are independent.

8.8 Bibliographical remarks and addenda

Kolmogorov's program

The general idea of on-line compression modeling seems to have originated in the work of Andrei Kolmogorov, who is perhaps best known for his axiomatization of probability theory as a branch of measure theory (Kolmogorov 1933a). Kolmogorov, however, never believed that his measure-theoretic axioms per se provide a satisfactory foundation for the applications of probability (as opposed to the mathematical theory of probability). Starting from Kolmogorov 1963 he embarked on a program of creating a better foundation. There is no complete published description of Kolmogorov's program, but his papers (Kolmogorov 1968, Kolmogorov 1983) and papers reporting work done by his PhD students (Martin-Löf 1966, Vovk 1986, Asarin 1987, 1988) provide material for a more or less plausible reconstruction of its main ideas; such an attempt was made in Vovk 2001b, Vovk and Shafer 2003.

The standard approach to modeling uncertainty is to choose a family of probability distributions, called a statistical model (see §A.1), one of which is believed to be the true distribution generating, or explaining in a satisfactory way, the data. (In some applications of probability theory, the true distribution is assumed to be known, and so the statistical model is a one-element set. In Bayesian statistics, the statistical model is complemented by another element, a prior probability distribution on the elements of the statistical model.) All modern applications of probability are widely believed to depend on this kind of modeling. (We saw in §8.5 that, e.g., even such a classical procedure as the confidence predictor (8.10) based on the t -distribution can be directly analyzed in terms of on-line compression models, but the standard approach would be to do analysis in terms of statistical models.)

In 1965–1970 Kolmogorov suggested a different approach to modeling uncertainty, based on information theory, with the purpose of providing a more direct link between the theory and applications of probability. He started, in Kolmogorov 1963, from the idea that the object of probability theory is a finitary version of von Mises's notion of collectives, which he called “tables of random numbers”, but the development of this idea lead him to the general idea of compression modeling. In Kolmogorov 1968 (§2) he replaced finitary collectives by “Bernoulli sequences”, which provide the first example of what we call “Kolmogorov complexity models”. This development became possible only after the introduction of the algorithmic notion of complexity (now called *Kolmogorov complexity*) in Kolmogorov 1965. In Kolmogorov 1983 he defines Markov binary sequences, another example. A third example, Gaussian sequences of real numbers, is described by Asarin (1987, 1988); Asarin 1987 also describes the Poisson model. We will describe these three examples after a brief general description of Kolmogorov's approach.

The general idea of Kolmogorov's program is that

practical conclusions of probability theory can be substantiated as implications of hypotheses of a *limiting*, under given constraints, complexity of phenomena under study

(Kolmogorov 1983, §4). In essence, a Kolmogorov complexity model is a way of summarizing information in a data sequence; the summary then provides the constraints under which the complexity of the data sequence is required to be close to the maximum. Using Kolmogorov's algorithmic notion of randomness (a data sequence x

is *algorithmically random* in a set A containing x if the Kolmogorov complexity $K(x|A)$ is close to the binary logarithm $\log |A|$), we can say that the data sequence is required to be algorithmically random given the summary (i.e., algorithmically random in the set of all data sequences with the same summary).

Each specific Kolmogorov complexity model provides a way of summarizing information in a data sequence. The Bernoulli and Markov models are for binary (consisting of 0 and 1) sequences. The Bernoulli model summarizes a binary sequence by the number of 1s in it. The Markov model summarizes it by the number of 1s after 1s, 1s after 0s, 0s after 1s, and 0s after 0s. Besides, it is always assumed that the length of the data sequence is part of the summary. Accordingly, a finite binary sequence is *Bernoulli* if it has a maximal Kolmogorov complexity in the set of binary sequences of the same length and the same number of 1s; *Markov* binary sequences have a maximal Kolmogorov complexity in the set of binary sequences with the same number of 1s after 1s, 1s after 0s, 0s after 1s, and 0s after 0s. The Gaussian model summarizes a sequence of real numbers by approximate values for its arithmetic mean and variance (8.6), and so *Gaussian* sequences of real numbers are those maximally complex in the set of sequences of the same length and with similar mean and variance.

The main features of Kolmogorov's program appear to be the following (some of these features have not been discussed so far):

- It is based on the idea of *compression*. The compact summary contains, intuitively, all useful information in the data.
- The idea that if the summary is known, the information left in the data is noise, is formalized using the *algorithmic* notion of Kolmogorov complexity: the complexity of the data under the constraint given by the summary should be maximal (the requirement of algorithmic randomness).
- Semantically, the requirement of algorithmic randomness means that the conditional distribution of the data given the summary is *uniform*.
- It is preferable to deduce properties of data sequences *directly* from the assumption of limiting complexity, without a detour through standard statistical models (examples of such direct inferences are given in Asarin 1987 and Asarin 1988 and hinted at in Kolmogorov 1983), especially that Kolmogorov complexity models are not completely equivalent to standard statistical models (Vovk 1986).
- Kolmogorov's program deals only with finite sets and their elements. This *finitary nature* of Kolmogorov's program is typical of Kolmogorov's work in general: e.g., in his 1928 and 1929 papers he found it helpful to state even such an apparently asymptotic result as the law of the iterated logarithm in the observable terms, for finite sequences.

Repetitive structures

The notion of repetitive structure, first introduced by Martin-Löf (1974), is a natural outgrowth of Kolmogorov's program. Different authors used the term “repetitive structure” in different senses (in this book we continue this tradition in that our notion of repetitive structure is somewhat different from those we have seen in literature), and so we will be using “repetitive structure” as a generic term covering several related concepts.

Martin-Löf spent 1964–1965 in Moscow as Kolmogorov's PhD student. His paper Martin-Löf 1966 is an important contribution to the study of the Bernoulli

model, but perhaps his main achievement (published in the same paper) in this area is to restate Kolmogorov's algorithmic notion of randomness in terms of universal statistical tests, thus demonstrating its fundamental character. After 1965 he and Kolmogorov worked on the information-theoretic approach to the applications of probability independently of each other (Martin-Löf having returned to Sweden), but arrived at similar concepts. In his work Martin-Löf was inspired not only by Kolmogorov's program but also by Gibbs's and Khinchin's ideas in statistical mechanics; this is the origin of names such as "Boltzmann distributions" or "microcanonical distributions" in the theory of repetitive structures.

Martin-Löf (1974) gave the definition of repetitive structure, as in §8.3, but with the conditional distributions $P_n(\cdot | \sigma)$ being uniform on a finite set and without the condition that t_n should be computable from t_{n-1} and z_n ; besides, his requirement of consistency of P_n involved conditioning on t_{n-1} only (and not on z_n).

Martin-Löf's theory of repetitive structures shared the idea of compression and the uniformity of conditional distributions with Kolmogorov's program; in fact the idea of compression was stated explicitly for the first time. An extra feature of repetitive structures is their *on-line character*: one consider sequences of all lengths n simultaneously (although the on-line character of the statistics t_n seems to have entered the theory of repetitive structures through the work of Steffen Lauritzen, who in Lauritzen 1988 attributes this notion to Freedman 1962, 1963).

Despite being a key contributor to the algorithmic theory of randomness, Martin-Löf did not use the algorithmic notions of complexity and randomness in his theory of repetitive structures. In Chap. 2 we already referred to our observation (Vovk and Shafer 2003) that these algorithmic notions tend not to lead to mathematical results in their strongest and most elegant form. After having performed their role as a valuable source of intuition, they are often discarded.

The notion of repetitive structure was later studied by Lauritzen; see, especially, his 1982 book and its revised and updated version (1988). Lauritzen's (1988, p. 207) repetitive structures do not involve any probabilities, which enter the picture through parametric "projective statistical fields".

Dawid (1982) was influential in propagating Martin-Löf and Lauritzen's ideas among Bayesian statisticians.

Freedman and Diaconis independently came up with ideas similar to Kolmogorov's (Freedman's first paper in this direction was published in 1962); they were inspired by de Finetti's theorem and the Krylov–Bogolyubov approach to ergodic theory.

The general theory of repetitive structures (as well as its most important special case, the exchangeability model) is now considered to be central to Bayesian model-building. See, e.g., the textbook by Bernardo and Smith (1994, Chap. 4), which discusses a wide range of repetitive structures, although without using this name.

The usual approach in the theory of repetitive structures is first to find all probability distributions P on \mathbb{Z}^∞ that agree with the given repetitive structure and then to take the extreme points of the set of all such P as the statistical model. (And once we have a statistical model, a wide arsenal of standard methods can be used.) In the next three subsections we will see that this strategy have been very successful in the case of the repetitive structures considered in this chapter, and then we will return to the general theory.

Exchangeability model

According to Hewitt and Savage 1955, Haag (1928) seems to be the first to discuss the concept of exchangeability; he also hinted at de Finetti's theorem but did not state it explicitly. De Finetti's theorem in the binary case was obtained in de Finetti 1930 and independently by Khinchin (1932); the extension to the general Borel case (stated as the result about real-valued random variables) is due to de Finetti (1937), and an abstract statement first appeared in Hewitt and Savage 1955.

As described in the previous subsection, the binary exchangeability ("Bernoulli") model was the first complexity model considered by Kolmogorov (who arrived at it developing von Mises's ideas).

There exists vast literature (including de Finetti 1938 and Freedman 1962) on partial exchangeability, a generalization of exchangeability akin to the Mondrian-exchangeability models.

Label-conditional OCMs were first described in print by Ryabko (2003).

Gaussian model

The origins of the Gaussian model lie in statistical physics. Especially important is the simplified version of the Gaussian model in which the sufficient statistics are

$$t_n(z_1, \dots, z_n) = \sqrt{z_1^2 + \dots + z_n^2},$$

and the conditional distributions $P_n(\cdot | \sigma)$ are uniform on the sphere $t_n = \sigma$. Gibbs proposed this as a model of the situation where z_i , $i = 1, \dots, n$, is the speed of the i th molecule of ideal gas; the value of t_n being known corresponds to the total energy of all molecules being known; the uniform conditional distributions are called micro-canonical distributions in statistical physics. Maxwell's law (obtained by Maxwell using the assumption of independence of the components of the molecules' speed along the Cartesian axes) states that the distribution of z_i is normal with zero mean and same variance. According to Bourbaki (1969), Borel (1914) appears to be the first to notice that Maxwell's law (representing the standard approach to statistical modeling) is a corollary of Gibbs's model when n is large. Borel's result was later developed by Gâteaux and Lévy (Lévy 1922).

The result that those probability distributions that agree with the simplified Gaussian model are mixtures of power normal distributions $N_{0,\sigma}^\infty$ appears to be due to Freedman (1963) (according to Kingman 1978), but it was also independently discovered by Kingman himself (Kingman 1972). The result about the full Gaussian model is due to Smith (1981).

Fisher was the first to prove that (8.9) has the t -distribution with $n - 2$ degrees of freedom (see Fisher 1922, pp. 610–611, Fisher 1925, §5, and Fisher 1973b). This result was explicitly used for the purpose of prediction by Baker (1935).

The name "Gauss liner model" was suggested for the standard linear regression model by Seal (1967). The on-line compression version of this model was suggested by Vovk (2004).

Markov model

Markov chains were introduced by Markov in 1906.

It was shown by Diaconis and Freedman (1980) that those probability distributions P that agree with the Markov on-line compression model (with \mathbf{Z} finite) are essentially mixtures of Markov-chain distributions (more accurately, each extreme distribution P is a recurrent Markov chain preceded by a random string of transient states with given transition counts; two strings of transient states having the same transition counts have the same probability).

The BEST theorem is named after its authors, de Bruijn, van Aardenne-Ehrenfest, Smith, and Tutte.

Kolmogorov's modeling vs. standard statistical modeling

The most important difference between Kolmogorov's program and the theory of repetitive structures seems to be in their basic goal: for the latter, it is derivation of standard statistical models, whereas Kolmogorov's main intention was to use complexity models directly. To achieve its goal, the theory of repetitive structures to a large degree abandoned Kolmogorov's finitary ideal: a brief glance at, e.g., Lauritzen 1988 reveals that the main action takes place at infinity.

Carrying over some properties of standard statistical models to complexity models was a subsidiary goal for Kolmogorov, but it is impossible to derive the standard models themselves in the absence of the on-line framework brought in by Martin-Löf and without making heavy use of the infinitary aspects of the framework. For example, papers by Asarin (1987, 1988) demonstrate that the Gaussian bell can be discerned in the Gaussian complexity model, but the corresponding mathematical results only have a limited accuracy, with full accuracy not achievable. This has important implications for confidence prediction under the two kinds of models, and we will discuss them in detail for the Bernoulli and Markov cases.

In the rest of this section we will assume that the reader is familiar with the main definitions of the algorithmic theory of randomness (see, e.g., Li and Vitányi 1997, V'yugin 1994, Vovk and V'yugin 1993) or is willing to trust the following intuitive picture. If P is a probability distribution on a measurable space Ω , we say that a function $t : \Omega \rightarrow [0, 1]$ is a P -test if, for any $\epsilon > 0$,

$$P\{\omega \in \Omega : t(\omega) \leq \epsilon\} \leq \epsilon.$$

Such functions t can serve as statistical tests for testing the hypothesis P (and were used as a technical tool in §5.3). We only consider *uniform tests*: functions $t_P(\omega)$ of two arguments, P (ranging over a wide class of probability distributions on a wide range of measurable spaces Ω ; it is important that Ω is not fixed in this definition) and ω (ranging over Ω), which are “upper semicomputable” (it suffices to know that this is the most natural notion of computability in this context) and for all P are P -tests as functions of ω . Martin-Löf's (1966) argument shows that there exists a *universal* uniform test, which is smaller, to within a multiplicative constant, than any other uniform test. (Although uniform tests were first formally introduced by Levin 1976.) We fix one of universal uniform tests, denote it by Δ , and call $\Delta_P(\omega)$ the *algorithmic randomness level* of ω w.r. to P . The *algorithmic randomness deficiency* $D_P(\omega)$ of $\omega \in \Omega$ w.r. to P is defined to be $-\log \Delta_P(\omega)$ (with \log being the base 2 logarithm), the algorithmic randomness level on the logarithmic scale.

If a set Ω is finite, $D_\Omega(\omega)$ stands for $D_P(\omega)$, where P is the uniform distribution on Ω ; Kolmogorov's original definition of $D_\Omega(\omega)$ was given in terms of Kolmogorov

complexity (as described on p. 215). If \mathcal{P} is a family of probability distributions on Ω ,

$$D_{\mathcal{P}}(\omega) := \inf_{P \in \mathcal{P}} D_P(\omega), \quad \omega \in \Omega. \quad (8.23)$$

(Levin 1973 showed that, for a wide range of \mathcal{P} , $D_{\mathcal{P}}$ can also be defined directly, generalizing the definition of D_P .) We say that $\omega \in \Omega$ is *algorithmically random*, or *typical*, w.r. to \mathcal{P} to mean that $D_{\mathcal{P}}(\omega)$ is small.

Let $\mathbf{Z} = \{0, 1\}$; we will consider two repetitive structures: Bernoulli and Markov. In the Bernoulli structure, the sufficient statistic $t_n(z_1, \dots, z_n)$ is the number of 1s among z_1, \dots, z_n , and in the Markov structure the components of the sufficient statistic $t_n(z_1, \dots, z_n)$ are the numbers of transitions

$$n_{i,j} := |\{k = 1, \dots, n-1 : z_k = i, z_{k+1} = j\}|, \quad i, j \in \{0, 1\},$$

and the initial bit z_1 . The conditional distributions $P_n(\cdot | \sigma)$ are always uniform.

When would we regard a binary sequence $\omega = (z_1, \dots, z_n)$ as a typical outcome generated by a repetitive structure (t_n) (Bernoulli or Markov)? There are two natural answers. We say that ω is *Kolmogorov-typical* if $D_{t_n^{-1}(\omega)}(\omega)$ (called the *Kolmogorov algorithmic randomness deficiency*, which we will abbreviate to “Kolmogorov deficiency”) is small. This notion was introduced in Kolmogorov 1968, 1983. We say that ω is *repetitive-typical* if $D_{\mathcal{P}}(\omega)$ (called the *repetitive algorithmic randomness deficiency* or, briefly, “repetitive deficiency”) is small, where \mathcal{P} is the class of all probability distributions on \mathbf{Z}^∞ that agree with the repetitive structure.

Both notions of typicalness can be applied to prediction. Namely, suppose we know the first $n-1$ examples z_1, \dots, z_{n-1} and our goal is to give a categorical prediction (0 or 1) for z_n , with a given repetitive structure accepted as the model for z_1, z_2, \dots . The prediction is possible if either $z_1, \dots, z_{n-1}, 0$ or $z_1, \dots, z_{n-1}, 1$ is untypical (we do not expect both to be untypical, since z_1, \dots, z_{n-1}, z_n is expected to be typical). We are interested in how different the two notions of typicalness are for the Bernoulli and Markov models and in the implications of any difference for prediction. (Our conclusion will be that there is an appreciable difference for both models, but it affects the quality of prediction significantly only for the Markov model.)

The case of the Bernoulli model is studied in Vovk 1986. It is shown that the requirement of repetitive typicalness is stronger; namely, $\omega = (z_1, \dots, z_n)$ will be repetitive-typical if and only if ω is Kolmogorov-typical and $t_n(\omega)$ (the number of 1s in ω) is typical w.r. to the binomial model. This statement carries over to a wide range of repetitive structures (including Markov). For each probability distribution P on \mathbf{Z}^∞ that agrees with the given repetitive structure, Pt_n^{-1} is the image of P under the mapping $(z_1, z_2, \dots) \mapsto t_n(z_1, \dots, z_n)$; let us denote the set of all such Pt_n^{-1} by $\mathcal{P}t_n^{-1}$. Then ω is repetitive-typical if and only if ω is Kolmogorov-typical and $t_n(\omega)$ is typical w.r. to $\mathcal{P}t_n^{-1}$ (in the sequel we will omit “w.r. to $\mathcal{P}t_n^{-1}$ ” when talking about the algorithmic randomness of $t_n(\omega)$).

It is now easy to see that there are Kolmogorov-typical sequences which are not repetitive-typical in the Bernoulli model: any Kolmogorov-typical sequence of a large even length containing precisely $n/2$ 1s will have repetitive deficiency of approximately $\frac{1}{2} \log n$. (The values of algorithmic randomness deficiency are only defined to the $O(1)$ accuracy and will always be approximate; we will not always mention this explicitly.) The value $\frac{1}{2} \log n$ is actually the largest possible value of repetitive deficiency for Kolmogorov-typical sequences of length n . It might appear

small as compared to the largest possible value for the Kolmogorov or repetitive deficiency, n , but it is comparable to the largest lower bound $\log n$ on the latter provided by the method of conformal prediction (remember that p-values used in conformal prediction are never less than $1/n$, which gives $\log n$ on the logarithmic scale; see the left-hand side of (2.18) on p. 26). Therefore, the difference of $\frac{1}{2} \log n$ might *a priori* have serious implications for prediction. The fact that it does not, follows from the “continuity” of binomial typicalness: the typicalness of the number of 1s seen cannot change much after observing one more example (this follows from the characterization of binomial typicalness in Vovk 1986 and is proved for the general exchangeability model in Nouretdinov et al. 2003). Since we expect z_1, \dots, z_{n-1} to be repetitive-typical when predicting z_n , $t_{n-1}(z_1, \dots, z_{n-1})$ will be typical; the property of continuity then implies that both $t_n(z_1, \dots, z_{n-1}, 0)$ and $t_n(z_1, \dots, z_{n-1}, 1)$ are typical, and in this case the difference between Kolmogorov typicalness and repetitive typicalness disappears.

The conformal method rejects a possible value, 0 or 1, for z_n based on the conditional distribution $P_n(\cdot | t_n)$; it does not take into account how untypical the value of the statistic t_n may become. Therefore, it can never work better than the idealized method based on Kolmogorov typicalness. This is not a problem for the Bernoulli (more generally, exchangeability) model; we will now see that it creates difficulties for the Markov model.

The key difference of the Markov model from the Bernoulli model is the absence of the continuity property for the algorithmic randomness deficiency of the summary $\sigma_n := t_n(z_1, \dots, z_n)$: it is possible that σ_n will become very untypical even for a typical σ_{n-1} . For concreteness, let us consider the following specific binary Markov chain. The initial state is 0 and the transition probabilities are: 0 is followed by 1 with probability M^{-2} and 1 is followed by 0 with probability M^{-1} , where M is a large positive number. Therefore, the usual state of the Markov chain is 0; we will expect typical runs of 0 to have the order of magnitude M^2 and typical runs of 1 to have the order of magnitude M .

Suppose the current state is $z_{n-1} = 0$ and n is very large; how unlikely is it that $z_n = 1$? The answer given by Kolmogorov typicalness will agree with the answer (8.17) (see p. 210) given by the conformal predictor but will be very different from the answer given by repetitive typicalness.

Let us start from repetitive typicalness. If z_1, \dots, z_{n-1} contains both subsequences 0, 1 and 1, 0 and is repetitive-typical w.r. to the Markov model, it will be typical w.r. to a Markov probability distribution P (see (8.23) and Diaconis and Freedman 1980, Example (19) on p. 120). The infimum in the definition (8.23) of the repetitive deficiency of z_1, \dots, z_n will be achieved at P or nearby (the universal uniform test will detect that the other transition probabilities disagree with the empirical data), and then the occurrence of the rare transition 0 → 1 at the end of z_1, \dots, z_n will raise the repetitive deficiency of such a sequence to $2 \log M$. Therefore, we expect $z_n = 0$ and the strength of this expectation is reflected, on the logarithmic scale, by the number $2 \log M$ (which corresponds to outputting the prediction set $\{0\}$ at the significance level M^{-2}).

The situation with conformal prediction appears anomalous: (8.17) involves the ratio $n_{1,0}/(n_{1,0} + n_{1,1})$ instead of something like $n_{0,1}/(n_{0,0} + n_{0,1})$. The former ratio gives probability M^{-1} , which is $\log M$ on the logarithmic scale. In the next paragraph we will see that this result is a consequence of using Kolmogorov’s strategy of evaluating the typicalness of $z_1, \dots, z_{n-1}, z'_n$ (where z'_n is a potential value for

the example z_n) given the value $t_n(z_1, \dots, z_{n-1}, z'_n)$; in particular, we will get the strength of expectation $\log M$ for $z_n = 0$ after observing $z_{n-1} = 0$ even using Kolmogorov deficiency (which is an ideal universal function, not computable in any practical sense, unlike the simple formula (8.17)).

Indeed, let $z_{n-1} = 0$, $z'_n = 1$ and suppose that we are given $\sigma_n := t_n(z_1, \dots, z_{n-1}, z'_n)$; in particular, knowing σ_n implies knowing that $z'_n = 1$. The uniform conditional probability (given σ_n) of $z_{n-1} = 1$ is about M times as large as the uniform conditional probability of $z_{n-1} = 0$; therefore, the unusual event that $z_1, \dots, z_{n-1}, z'_n$ ends in 01 raises the Kolmogorov deficiency of the sequence to $\log M$, which is $\log M$ short of the repetitive deficiency. It is easy to see where the $\log M$ is lost. The sequence $z_1, \dots, z_{n-1}, z'_n$ is untypical for two reasons: first, σ_n is untypical (the algorithmic randomness deficiency of σ_n is $\log M$, since having 1 as the last element is about M times less likely than having 0 as the last element); second, even given σ_n , the sequence $z_1, \dots, z_{n-1}, z'_n$ is still untypical (another $\log M$). Repetitive typicalness takes both sources into account, whereas Kolmogorov typicalness disregards the first source.

We can see that there is little hope of obtaining for Markov chains, even binary, a result about the optimality of conformal predictors analogous to that of Chap. 3 (see Theorem 3.1 and the construction of a universal confidence predictor), unless symmetry (equality of the transition probabilities for $0 \rightarrow 1$ and $1 \rightarrow 0$) is assumed. There are several ways to improve the efficiency of conformal prediction for asymmetric Markov chains. We could consider a “team” of two conformal predictors, one predicting the examples following 0, and the other predicting the examples following 1, as in §7.3. It might be possible to use Kolmogorov’s (1983) suggestion to summarize the number of transitions *approximately*. We could use a version of conformal predictor for predicting several examples (this possibility will also be discussed in Chap. 10), but only report the prediction for the next example. Finally, we could simply estimate the transition probabilities, as described at the end of §10.1. We are not, however, especially interested in the fixes for the specific case of asymmetric Markov chains. Our goal in this discussion has been to attract the reader’s attention to the necessity of exercising care in the use of on-line compression models.

On-line compression modeling II: Venn prediction

This chapter extends the notion of a Venn predictor (Chap. 6) to the general framework of on-line compression modeling (Chap. 8). The result stated in Chap. 6 that Venn predictors are automatically valid is extended from the exchangeability model to general on-line compression models; its proof, given in §9.6, also proves the result of Chap. 6.

Another focus of this chapter is introduction of a new class of on-line compression models, which we call hypergraphical models; it is analogous to causal networks in machine learning and contingency tables in mathematical statistics. Venn predictors appear to be particularly suitable for this model, although we also briefly mention conformal prediction. For each hypergraphical model we define the most refined Venn predictor, which we call the “fully conditional Venn predictor” (FCVP).

A special attention is paid to a subclass of hypergraphical models consisting of what we call “junction-tree models”. This subclass is both amenable to straightforward analysis and wide enough to cover many practically interesting models. In particular, we find a simple explicit representation of the fully conditional Venn predictor for junction-tree models. In §9.5 we demonstrate the working of the fully conditional Venn predictor on an artificial data set randomly generated from a simple hypergraphical model. It appears that the FCVP will give reasonable results for small models, but to deal with larger models one would have to abandon full conditioning.

Historically, the main sources of hypergraphical models are statistical physics, path analysis, and contingency tables (Lauritzen 1996, p. 1), but the area providing most of the hypergraphical models of interest in our present context is the theory of causal networks; the model of §9.5 is of this type. Transition to junction-tree models is the standard procedure in that area (Jensen 1996, Shafer 1996b, Cowell et al. 1999).

In this chapter we only consider the problem of classification: the label space \mathbf{Y} is finite.

9.1 Venn prediction in on-line compression models

The definition of validity for multiprobability predictors given in §6.2 is generalized to arbitrary on-line compression models (or N -models, as defined on p. 194) $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ as follows: a multiprobability predictor F is N -valid, where N is a positive integer (the horizon), if for any probability distribution P on \mathbf{Z}^N that agrees with M and any nonnegative reversible game supermartingale G with horizon N and $G(\square) = 1$, there exists a P -supermartingale S_0, \dots, S_N with $S_0 = 1$ such that (6.19) (p. 156) holds for all $x_1, y_1, \dots, x_N, y_N$.

Next we generalize the notion of Venn predictor, introduced in § 6.3 for the exchangeability model, to arbitrary on-line compression models. Any sequence of measurable finite partitions A_n of the sets $\Sigma_{n-1} \times \mathbf{Z}$, $n = 1, 2, \dots$, is called a *taxonomy* (or *Venn taxonomy*); as always, $A_n(\sigma, z)$ stands for the element of the partition A_n that contains (σ, z) .

The *Venn predictor determined by* (A_n) is the multiprobability predictor which outputs $P_n := \{p_y : y \in \mathbf{Y}\} \subseteq \mathbf{P}(\mathbf{Y})$ at the n th trial, where each probability distribution p_y on \mathbf{Y} is defined as follows. Complement the new object x_n by the “postulated label” y and set

$$\begin{aligned}\sigma_{n-1} &:= t_{n-1}(z_1, \dots, z_{n-1}), \\ \sigma_n &:= t_n(z_1, \dots, z_{n-1}, (x_n, y)).\end{aligned}$$

Define p_y to be the probability distribution under $B_n(\sigma_n)$ of the labels in $A_n(\sigma_{n-1}, (x_n, y))$: for all $y' \in \mathbf{Y}$,

$$p_y\{y'\} := \frac{B_n(\{(\sigma_{n-1}^*, (x^*, y^*)) \in A_n(\sigma_{n-1}, (x_n, y)) : y^* = y'\} | \sigma_n)}{B_n(A_n(\sigma_{n-1}, (x_n, y)) | \sigma_n)}. \quad (9.1)$$

A *Venn predictor* is the Venn predictor determined by some taxonomy.

Formula (9.1) can be spelled out in the same way as we did for the special case of the exchangeability model in §6.3: partition the set $F_n^{-1}(\sigma_n) \subseteq \Sigma_{n-1} \times \mathbf{Z}$ into categories, assigning (σ', z') and (σ'', z'') to the same category if and only if $A_n(\sigma', z') = A_n(\sigma'', z'')$, and define p_y to be the probability distribution of the labels in the category T containing $(\sigma_{n-1}, (x_n, y))$:

$$p_y\{y'\} := \frac{B_n(\{(\sigma^*, (x^*, y^*)) \in T : y^* = y'\} | \sigma_n)}{B_n(T | \sigma_n)}.$$

Theorem 9.1. *Any Venn predictor is N -valid for any N .*

The proof of this theorem is given in §9.6.

9.2 Generality of finitary repetitive structures

In this section we specialize the discussion of generality and specificity of repetitive structures in §8.4 (p. 197) to the Kolmogorov-type finitary repetitive structures with uniform conditional distributions. Formally, a repetitive

structure $(\Sigma, \mathbf{Z}, (t_n), (P_n))$ is *finitary* if the example space \mathbf{Z} is finite and, for all n and $\sigma \in t_n(\mathbf{Z}^n)$, $P_n(\cdot | \sigma)$ is the uniform probability distribution on the finite set $t_n^{-1}(\sigma)$. Let us fix \mathbf{Z} and Σ (the latter can be, e.g., chosen large enough to contain all summaries that we are likely to be interested in). Then a finitary repetitive structure is determined by (t_n) , and we will sometimes say that (t_n) is the repetitive structure.

The choice of the repetitive structure (t_n) reflects the strength of the assumption that we are willing to make about Reality. According to the definition given on p. 197, a finitary repetitive structure (t_n) is *more specific* than another finitary repetitive structure (t'_n) (denoted as $(t_n) \preceq (t'_n)$) if $t_n = f_n(t'_n)$ for some measurable functions $f_n : \Sigma \rightarrow \Sigma$. Intuitively, in this case t_n performs a greater data compression than t'_n does and so represents a stronger assumption about Reality. In particular, if $(t_n) \preceq (t'_n)$, then any probability distribution on \mathbf{Z}^∞ that agrees with (t_n) also agrees with (t'_n) . The analogous statement is also true for a finite horizon N : any probability distribution on \mathbf{Z}^N that agrees with (t_n) also agrees with (t'_n) .

We saw in Chap. 8 that the Gaussian model and the Mondrian models are more specific than the exchangeability model, and the partial order induced by the relation “more specific” on the Mondrian models is fairly rich. We will really need this relation, however, only in the case of hypergraphical models, introduced in the next section.

9.3 Hypergraphical models

Starting from this section we assume that the examples are structured, consisting of “variables”. Formally, a *hypergraphical structure* is a triple (V, \mathcal{E}, Ξ) where:

- V is a finite set whose elements will be called *variables*;
- \mathcal{E} is a family of V ’s subsets; elements of \mathcal{E} are called *clusters*; the union of all clusters is required to be the whole of V ;
- Ξ is a function that maps each variable $v \in V$ into a finite set $\Xi(v)$ of the “values that v can take”; $\Xi(v)$ is called the *frame* of v ; to exclude trivial cases, we always assume $\forall v \in V : |\Xi(v)| > 1$.

We will eventually assume that some of the variables are marked as *labels*, but this assumption will not be needed in many of our considerations. A *configuration* on a cluster E (or, more generally, V ’s subset) E is an assignment of an element of $\Xi(v)$ to each $v \in E$. An *example* is a configuration on V ; we take \mathbf{Z} to be the set of all examples.

A *table* on a cluster E is an assignment of a nonnegative number to each configuration on E . We will mainly be interested in *natural tables*, which assign only natural (i.e., nonnegative integer) numbers to configurations. (These are known as “contingency tables” in statistics.) The *size* of the table is the sum of values that it assigns to different configurations. A *table set* f assigns to

each cluster E a table f_E on this cluster. *Natural table sets* are table sets all of whose tables are natural. We will only be interested in table sets all of whose tables have the same size, which is then called the size of the table set. The number assigned by a natural table set σ to a configuration of a cluster E will sometimes be called the σ -count of that configuration.

Hypergraphical repetitive structures

Now we are ready to define the hypergraphical repetitive structure and OCM associated with a hypergraphical structure (V, \mathcal{E}, Ξ) ; as usual, we start from the repetitive structure $(\Sigma, \mathbf{Z}, (t_n), (P_n))$. The table set $t_n(z_1, \dots, z_n)$ generated by a data sequence (i.e., sequence of examples) z_1, \dots, z_n assigns to each configuration on each cluster the number of examples among z_1, \dots, z_n that agree with that configuration (we say that an example z *agrees* with a configuration on a cluster E if that configuration coincides with the restriction $z|_E$ of z to E). The number of data sequences generating a table set σ will be denoted $\#\sigma$ (for $\#\sigma$ to be non-zero the size of σ must exist, and then the length of each sequence generating σ will be equal to its size). The table sets σ with $\#\sigma > 0$ (called *consistent* table sets) are called *summaries*; they form the summary space Σ of the hypergraphical on-line compression model and repetitive structure associated with (V, \mathcal{E}, Ξ) . The conditional probability distribution $P_n(\cdot | \sigma)$, where n is the size of σ , is the uniform distribution on the set of all data sequences z_1, \dots, z_n that generate σ .

The explicit definition of the hypergraphical OCM $(\Sigma, \square, \mathbf{Z}, F, B)$ is as follows:

- Σ is the set of all summaries (i.e., consistent table sets); \square is the *empty* table set, i.e., the one of size 0;
- \mathbf{Z} is the set of all examples (i.e., configurations on V);
- the table set $F(\sigma, z)$ is obtained from σ by adding 1 to the σ -count of each configuration that agrees with z ;
- an example z *agrees* with a summary σ if the σ -count of each configuration that agrees with z is positive; if so, we obtain a table set denoted $\sigma \downarrow z$ from σ by subtracting 1 from the σ -count of any configuration that agrees with z ; $B_n(\sigma)$, where n is the size of σ , is defined by

$$B_n(\{(\sigma \downarrow z, z)\} | \sigma) := \frac{\#(\sigma \downarrow z)}{\#\sigma}.$$

Among the probability distributions P that agree with the hypergraphical structure (V, \mathcal{E}, Ξ) (i.e., with the OCM associated with (V, \mathcal{E}, Ξ) ; we do not always distinguish between hypergraphical structures and the corresponding OCMs and repetitive structures) are power distributions Q^∞ such that each Q (a probability distribution on \mathbf{Z}) decomposes into

$$Q\{a\} = \prod_{E \in \mathcal{E}} f_E(a|_E), \quad (9.2)$$

where a is any configuration on V , f is a fixed table set (not necessarily natural), and $a|_E$ is, as usual, the restriction of a to E .

The exchangeability model with the example space \mathbf{Z} corresponds to the hypergraphical model with only one cluster, $\mathcal{E} = \{V\}$.

Fully conditional Venn predictor

In the case of hypergraphical OCM with one or more vertices marked as labels, there exists a very natural Venn predictor, which will be called the *fully conditional Venn predictor* (FCVP). It is defined to be the Venn predictor determined by the taxonomy (called the *fully conditional taxonomy*) A_n in which $A_n(\sigma, z)$ consists of all (σ', z') for which z and z' coincide on all non-label variables.

The FCVP is not only natural but also computationally efficient in the class of hypergraphical models known as junction-tree models, introduced in the next section. It will be the only predictor considered in this chapter, but this should not be interpreted as a recommendation to always use it for hypergraphical models: carefully crafted Venn taxonomies will definitely have advantages for small data sets, and conformal predictors also remain an attractive option.

Generality of hypergraphical models

Fix the set V of variables and the frame $\Xi(v)$ for each variable. The following proposition answers the question when the repetitive structure corresponding to a cluster set \mathcal{E}_1 is more specific than the repetitive structure corresponding to a cluster set \mathcal{E}_2 (in this case we will say that \mathcal{E}_1 is more specific than \mathcal{E}_2).

Proposition 9.2. *A cluster set \mathcal{E}_1 is more specific than a cluster set \mathcal{E}_2 , denoted $\mathcal{E}_1 \preceq \mathcal{E}_2$, if and only if for all $E_1 \in \mathcal{E}_1$ there exists $E_2 \in \mathcal{E}_2$ such that $E_1 \subseteq E_2$.*

Proof. The part “if” is obvious, so we only prove “only if”. Suppose \mathcal{E}_1 is more specific than \mathcal{E}_2 but there exists an $E \in \mathcal{E}_1$ which is not covered by any element of \mathcal{E}_2 . Let $k := |E|$; without loss of generality we suppose that the model is binary ($\Xi(v) = \{0, 1\}$ for all v), that $E = V$, and that all subsets of E of size $k - 1$ are in \mathcal{E}_2 .

Consider the following $k2^{k-1} \times 2^k$ matrix X : the columns of X are indexed by $\{0, 1\}^k$ (they represent the configurations of E); the rows of X are indexed by the sequences in $\{0, 1\}^k$ in which one of the bits is replaced by the symbol “x” (they represent the configurations of the k subsets of E of size $k - 1$); the element $X_{i,j}$ of X in row i and column j is 1 if j can be obtained from i by replacing the x with 0 or 1 and is 0 otherwise. If the table corresponding to the cluster E is given by a vector $t \in \mathbb{N}_0^{2^k}$ (\mathbb{N}_0 being the set $\{0, 1, \dots\}$ of nonnegative integers), the tables corresponding to the subsets of E of size

$k - 1$ are given by Xt . Therefore, our goal will be achieved if we show that there are two vectors $t_1, t_2 \in \mathbb{N}_0^{2^k}$ such that $Xt_1 = Xt_2$.

The rank of the matrix X is at most $2^k - 1$, since the sum of all columns is the identical 2. The procedure of Gaussian elimination shows that there exists a zero linear combination of X 's columns with rational (and, therefore, with integer) coefficients. Let t_0 be a vector in \mathbb{Z}^{2^k} such that $Xt_0 = 0$. Now we can take as t_1 any vector in \mathbb{N}^{2^k} with sufficiently large elements and set $t_2 := t_1 + t_0$. \square

For simplicity we will only consider *reduced* hypergraphical models, i.e., models (V, \mathcal{E}, Ξ) such that no $E_1, E_2 \in \mathcal{E}$ are nested, $E_1 \subseteq E_2$; in this case we will also say that \mathcal{E} is reduced. This does not limit generality, since we can always replace \mathcal{E} by $\text{red}(\mathcal{E})$, where $\text{red}(\mathcal{E})$ is \mathcal{E} with all $E \in \mathcal{E}$ contained in some $E' \in \mathcal{E}$ removed. (In other words, $\text{red}(\mathcal{E})$ is the only reduced element of \mathcal{E} 's equivalence class, where the equivalence of \mathcal{E}_1 and \mathcal{E}_2 means that $\mathcal{E}_1 \preceq \mathcal{E}_2$ and $\mathcal{E}_2 \preceq \mathcal{E}_1$.)

The set of all reduced hypergraphical structures \mathcal{E} (with V and Ξ fixed) forms a lattice with the join and meet operations

$$\begin{aligned}\mathcal{E}_1 \vee \mathcal{E}_2 &:= \text{red}(\mathcal{E}_1 \cup \mathcal{E}_2), \\ \mathcal{E}_1 \wedge \mathcal{E}_2 &:= \text{red}\{E_1 \cap E_2 : E_1 \in \mathcal{E}_1, E_2 \in \mathcal{E}_2\}.\end{aligned}$$

9.4 Junction-tree models

An important special case is where we can arrange the clusters of a hypergraphical model into a “junction tree”. We will be able to give efficient prediction algorithms only for such junction-tree models; if the hypergraphical model we happen to be interested in is not of this type, it should be replaced by a more general junction-tree model before our prediction algorithms can be applied.

Formally, a *junction tree* for a hypergraphical model (V, \mathcal{E}, Ξ) is an undirected tree (U, S) (with U the set of vertices and S the set of edges) together with a bijective mapping C from the vertices U of the tree to the clusters \mathcal{E} of the hypergraphical model which satisfies the following property: if a vertex v lies on the path from a vertex u to a vertex w in the tree (U, S) , then

$$C_u \cap C_w \subseteq C_v$$

(we let C_x stand for $C(x)$). The tree (U, S) will also sometimes be called the junction tree (when the bijection is clear from the context). It is convenient to identify vertices v of the junction tree with the corresponding clusters C_v in \mathcal{E} . If $s = \{u, v\} \in S$ is an edge of the junction tree connecting vertices u and v , we will write C_s for $C_u \cap C_v$; C_s will be called the *separator* between C_u and C_v .

We will say “junction-tree structures/models” to mean hypergraphical structures/models in which the clusters are arranged into a junction tree. Fix such a model (V, \mathcal{E}, Ξ) until the end of this section; (U, S) is the corresponding junction tree, F_n are the forward functions, B_n are the backward kernels, t_n are the statistics, and P_n are the conditional probability distributions given a summary.

Combinatorics of junction-tree models

It is easy to characterize consistent table sets in junction-tree structures. If $E_1 \subseteq E_2 \subseteq V$ and f is a table on E_2 , its *marginalization* to E_1 is the table f^* on E_1 such that $f^*(a) = \sum_b f(b)$ for all configurations a on E_1 , where b ranges over all configurations on E_2 that agree with a (i.e., such that $b|_{E_1} = a$).

Lemma 9.3. *A natural table set σ on (V, \mathcal{E}, Ξ) is consistent if and only if the following two conditions hold:*

- *each table in σ is of the same size;*
- *if clusters $E_1, E_2 \in \mathcal{E}$ intersect, the marginalizations of their tables to $E_1 \cap E_2$ coincide.*

This lemma is obvious; it, however, ceases to be true if the assumption that (V, \mathcal{E}, Ξ) is a junction-tree structure is dropped.

If σ is a summary and E is a cluster, we earlier defined σ_E as the table that σ assigns to E . If E is a separator, say $E = C_{\{u,v\}}$, σ_E stands for the marginalization of σ_{C_u} (equivalently, by Lemma 9.3, of σ_{C_v}) to E .

The *factorial-product* of a cluster or separator E in a summary σ is, by definition,

$$\text{fp}_\sigma(E) := \prod_{a \in \text{conf}(E)} \sigma_E(a)!$$

(remember that $0! = 1$), where $\text{conf}(E)$ is the set of all configurations on E .

Lemma 9.4. *Consider a summary σ of size n in the junction-tree model. The number of data sequences of length n generating the table set σ equals*

$$\#\sigma = \frac{n! \prod_{s \in S} \text{fp}_\sigma(C_s)}{\prod_{u \in U} \text{fp}_\sigma(C_u)}. \quad (9.3)$$

Proof. The proof is by induction on the size of the junction tree. If the junction tree consists of only one vertex u , the right-hand side of (9.3) becomes

$$\frac{n!}{\text{fp}_\sigma(C_u)} = \frac{n!}{\prod_{a \in \text{conf}(C_u)} \sigma_{C_u}(a)!},$$

which is the correct multinomial coefficient.

Now let us assume that (9.3) is true for some tree and prove that it remains true for that tree extended by adding an edge s and a vertex u . (The example

space for the new tree will be bigger.) We are required to show that the number of data sequences generating σ is multiplied by

$$\frac{\text{fp}_\sigma(C_s)}{\text{fp}_\sigma(C_u)} = \prod_{a \in \text{conf}(C_s)} \frac{\sigma_{C_s}(a)!}{\prod_{b \in \text{agr}(a)} \sigma_{C_u}(b)!}, \quad (9.4)$$

where $\text{agr}(a)$ is the set of all configurations on C_u that agree with a . It remains to notice that the number of ways in which each sequence of n examples in the old tree can be extended to a sequence of n examples in the new tree is given by the right-hand side of (9.4). \square

We will use the shorthand

$$\sigma_u(z) := \sigma_{C_u}(z|_{C_u}),$$

for both vertices and edges u of a junction tree; here z is an example or, more generally, a configuration whose domain includes all variables in C_u .

Lemma 9.5. *Given the summary σ of the first n examples z_1, \dots, z_n , the $B_n(\sigma)$ -probability that $z_n = a$ equals (the maximum-likelihood estimate – see §9.6)*

$$\frac{\prod_{u \in U} \sigma_u(a)}{n \prod_{s \in S} \sigma_s(a)} \quad (9.5)$$

(this ratio is set to 0 if any of the factors in the numerator or denominator is 0; in this case $z_n = a$ does not agree with the summary σ).

Proof. Using Lemma 9.4, we obtain for the probability of $z_n = a$:

$$\frac{\#(\sigma \downarrow a)}{\#\sigma} = \frac{(n-1)! \prod_{s \in S} \text{fp}_{\sigma \downarrow a}(C_s) \prod_{u \in U} \text{fp}_\sigma(C_u)}{\prod_{u \in U} \text{fp}_{\sigma \downarrow a}(C_u) n! \prod_{s \in S} \text{fp}_\sigma(C_s)} = \frac{\prod_{u \in U} \sigma_u(a)}{n \prod_{s \in S} \sigma_s(a)}. \quad \square$$

Shuffling data sets

In this subsection we will see that Lemma 9.5 provides an efficient means of drawing a data set z_1, \dots, z_n from the conditional distribution $P_n(\sigma_n)$, where σ_n is a summary of size n in the junction-tree model. This can be used for shuffling data sets to make them conform to the given hypergraphical model (see §B.4).

It is convenient first to direct the junction tree, designating an arbitrary vertex as the root \square and directing all edges from the root (so that the root becomes an ancestor of every vertex). We can then rewrite (9.5) as

$$\frac{\sigma_\square(a)}{n} \prod_{u \in U \setminus \{\square\}} \frac{\sigma_u(a)}{\sigma_{u'}(a)}, \quad (9.6)$$

where u' is the separator between u and u 's parent. The last formula provides an efficient means of generating a random data sequence z_1, \dots, z_n from a

summary σ_n of size n : to generate z_n , first generate $z_n|_{C_\square}$ from σ_\square/n (i.e., from the probability distribution that assigns weight $\sigma_\square(a)/n$ to each configuration a on C_\square), then choose \square 's child u and generate $z_n|_{C_u}$ from $\sigma_u/\sigma_{u'}$ (i.e., from the probability distribution that assigns weight $\sigma_u(a)/\sigma_{u'}(a)$ to each configuration a on C_u that agrees with $z_n|_{u'}$), and so on. After z_n is generated, we can generate z_{n-1} in a similar way from $\sigma_n \downarrow z_n$, then generate z_{n-2} from $\sigma_n \downarrow z_n \downarrow z_{n-1}$, etc.

Decomposability in junction-tree models

Corollary 9.6. *Each power probability distribution Q^∞ on \mathbf{Z}^∞ that agrees with the junction-tree model is decomposable in the sense of (9.2).*

Proof. The idea of derivation of this corollary from Lemma 9.5 is standard; see, e.g., Lauritzen 1988 (Theorem 4.4 on p. 61).

Let Q be a probability distribution on \mathbf{Z} such that the power distribution Q^∞ on the set of sequences z_1, z_2, \dots agrees with the junction-tree model. It is clear that, for each configuration a on V ,

$$Q_n(a) := Q^\infty(z_1 = a \mid t_n(z_1, \dots, z_n), z_{n+1}, z_{n+2}, \dots)$$

(the conditional Q^∞ -probability that $z_1 = a$) is given by (9.6). According to Lévy's "downward" theorem (Williams 1991, §14.4) and the Hewitt–Savage zero-one law (Shiryayev 1996, Theorem IV.1.3), $Q_n(a)$ converge almost surely to $\mathbb{E}_{Q^\infty} Q_n(a) = Q\{a\}$ as $n \rightarrow \infty$. Borel's strong law of large numbers shows that all ratios in (9.6) converge almost surely; this completes the proof. \square

Prediction in junction-tree models

Next we consider the situation where the $|V|$ variables of the junction-tree model are divided into the *attributes* V_{obj} and the *label variables* V_{lab} ; only a subset of labels, the *target label variables* V_{targ} , have to be predicted¹. Therefore,

$$V_{\text{targ}} \subseteq V_{\text{lab}} = V \setminus V_{\text{obj}}.$$

The *object space* \mathbf{X} is then the set of all configurations on V_{obj} , the *label space* \mathbf{Y} is the set of all configurations on V_{lab} , and we define the *target label space* \mathbf{Y}^{targ} to be the set of all configurations on V_{targ} . Each example z_n has two components: the values $x_n := z_n|_{V_{\text{obj}}} \in \mathbf{X}$ (the *object*) taken by the attributes and the values $y_n := z_n|_{V_{\text{lab}}} \in \mathbf{Y}$ (the *label*) taken by the label variables; we will write (x_n, y_n) to mean z_n . The values $y_n^{\text{targ}} := z_n|_{V_{\text{targ}}} \in \mathbf{Y}^{\text{targ}}$ taken by the target label variables will be called the *target label*.

¹It might have been more natural to restrict the use of the word "label" only to target labels, and to call non-target labels, for example, nuisance variables; our terminology, however, is more consistent with that of the previous chapters.

In this subsection we will give a simple and explicit representation of the FCVP. Suppose we have observed examples z_1, \dots, z_{n-1} and we are given a new object x_n . For each $y'' \in \mathbf{Y}^{\text{targ}}$ we are interested in the conditional $B_n(\sigma_n)$ -probability (where $\sigma_n := t_n(z_1, \dots, z_{n-1}, (x_n, y))$, for different $y \in \mathbf{Y}$) that the target label y_n^{targ} is y'' given that the values of the attributes are x_n .

Consider the “ y -completion”, in which the object x_n is complemented to the example (x_n, y) , where $y \in \mathbf{Y}$ is a label. Let $A_{y,y'}$ be the following $|\mathbf{Y}| \times |\mathbf{Y}|$ matrix with rows and columns indexed by \mathbf{Y} : each entry $A_{y,y'}$ is the fraction of examples in $A_n(\sigma_{n-1}, (x_n, y))$ (where A_n is the full conditional taxonomy and $\sigma_{n-1} := t_{n-1}(z_1, \dots, z_{n-1})$) labeled by y' ; remember that this matrix determines the Venn predictor’s output (which is the set of the probability distributions represented by the rows of the matrix; cf. p. 161). This fraction (the conditional probability in the y -completion that $z_n = (x_n, y')$ given that the attributes’ values are x_n) is proportional to

$$\frac{\prod_{u \in U} \sigma_u((x_n, y'))}{\prod_{s \in S} \sigma_s((x_n, y'))} \quad (9.7)$$

where $\sigma := t_n(z_1, \dots, z_{n-1}, (x_n, y))$, since, by Lemma 9.5, (9.7) is proportional to the unconditional $B_n(\sigma)$ -probability that $z_n = (x_n, y')$. To obtain the prediction for the target labels only, the rows of A have to be marginalized to the target labels. Summarizing, we obtain the following description of the FCVP for junction-tree models (some explanations are given after the description).

JUNCTION-TREE FCVP

```

 $\sigma_0 := \square;$ 
FOR  $n = 1, 2, \dots$ :
  read  $x_n \in \mathbf{X}$ ;
  FOR  $y \in \mathbf{Y}$ :
     $\sigma := F_n(\sigma_{n-1}, (x_n, y));$ 
    FOR  $y' \in \mathbf{Y}$ 
       $A_{y,y'} := \frac{\prod_{u \in U} \sigma_u((x_n, y'))}{\prod_{s \in S} \sigma_s((x_n, y'))};$ 
    END FOR;
  END FOR;
  normalize the rows of  $A_{y,y'}$ ;
  set  $P_n \subseteq \mathbf{P}(\mathbf{Y}^{\text{targ}})$  to the rows of  $A_{y,y'}$  marginalized to  $V_{\text{targ}}$ ;
  read  $y_n \in \mathbf{Y}$ ;
   $\sigma_n := F_n(\sigma_{n-1}, (x_n, y_n))$ 
END FOR.
```

The normalization of the rows of the matrix $A_{y,y'}$ means that first the sums $S_y := \sum_{y'} A_{y,y'}$ are computed and then each $A_{y,y'}$ is divided by S_y . If $a(y')$, $y' \in \mathbf{Y}$, is a row of the matrix A , its marginalization to V_{targ} is defined to be the probability distribution on \mathbf{Y}^{targ} that gives the weight

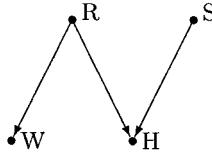


Fig. 9.1. The “wet grass” causal network

$$\sum_{y': y' | v_{\text{targ}} = y''} a(y')$$

to each $y'' \in \mathbf{Y}^{\text{targ}}$.

Universality of the fully conditional Venn predictor

The FCVP is universal in our usual asymptotic sense: if the examples are generated from a power probability distribution Q^∞ on \mathbf{Z}^∞ that agrees with the given junction-tree model, the maximum distance between the prediction it outputs and the true conditional probability distribution for the next label tends to zero (the *maximum distance* between a set A and a point b being defined as the supremum of distances between $a \in A$ and b). This follows from, e.g., the proof of Corollary 9.6.

9.5 Causal networks and a simple experiment

A rich source of hypergraphical models is provided by the theory of causal modeling. We start from a simple example (considered by Pearl 1988 and Jensen 1996).

One morning when Mr. Holmes leaves his house for work he notices that his grass is wet (H). Was there a rain overnight (R) or did he forget to turn off the sprinkler (S)? Next he checks his neighbor Dr. Watson’s grass (W). It is wet as well, and so Mr. Holmes concludes that wet grass was caused by rain. We can arrange the variables H (Mr. Holmes’s grass is wet), W (Dr. Watson’s grass is wet), R (rain), and S (sprinkler) in the causal network shown in Fig. 9.1. The directions of the arrows are intended to reflect the order in which the values of the variables are settled by Reality: first she decides, using a stochastic procedure, on the values of R and S ; given the realized values of R she decides on W ; and finally, given R and S she decides on H . The probabilities used by Reality are as follows (each variable is assumed binary and takes value 1 if the corresponding event happens and 0 if not):

$$\mathbb{P}\{R = 1\} = 0.2, \quad \mathbb{P}\{S = 1\} = 0.1 \tag{9.8}$$

are the probabilities for R and S ,

$$\mathbb{P}\{W = 1 | R = 1\} = 1, \quad \mathbb{P}\{W = 1 | R = 0\} = 0.2 \quad (9.9)$$

are the conditional probabilities for W given R (Watson may also forget to turn off his sprinkler, but this event is not reflected in the network explicitly), and

$$\begin{aligned} \mathbb{P}\{H = 1 | R = 1, S = 1\} &= 1, \quad \mathbb{P}\{H = 1 | R = 1, S = 0\} = 1, \\ \mathbb{P}\{H = 1 | R = 0, S = 1\} &= 0.9, \quad \mathbb{P}\{H = 1 | R = 0, S = 0\} = 0 \end{aligned} \quad (9.10)$$

are the conditional probabilities for H given R and S .

The general approach of causal modeling is to start from a directed acyclic graph (or, more generally, a “chain graph”), such as that in Fig. 9.1, erase the directions of all arrows adding undirected edges between all pairs of vertices that share a common child (“marrying the parents”), and then consider the hypergraphical model whose clusters are the cliques of the resulting undirected graph. If this model is not a junction-tree model (it is in the case of Fig. 9.1), there are computationally efficient ways to find its reasonable junction-tree extension (in the sense of the relation \preceq ; see Proposition 9.2 on p. 227). The area of causal modeling thus provides numerous examples of junction-tree models; for details, see Jensen 1996 or Cowell et al. 1999.

The preceding paragraph describes only part of the process of causal modeling, which is sometimes called qualitative modeling. Another important ingredient is quantitative modeling (Cowell et al. 1999, pp. 27–29): the standard approach requires both the structure (such as Fig. 9.1) and the prior probabilities (such as (9.8)–(9.10)). In our approach we do not need the second ingredient: given only the structure, Venn predictors output multiprobability predictions that are automatically valid.

The hypergraph corresponding to the “wet grass” network of Fig. 9.1 has two clusters, $\{W, R\}$ and $\{H, R, S\}$; they form a trivial junction tree with two vertices and one edge between them. The separator is $\{R\}$.

Table 9.1 shows the output of the FCVP when run on a data set generated randomly from (9.8)–(9.10). Mr. Holmes observes his own and Dr. Watson’s grass every morning, then checks his sprinkler (if in any doubt), and finally listens to the weather report in the car on his way to work (of course, this pure on-line protocol can be relaxed, as in Chap. 4). Table 9.1 gives, for selected trials, the trial number n , the observed values H_n , W_n , and S_n of the variables H , W , and S at trial n , the prediction (more precisely, the convex hull of the computed multiprobability) for $S_n = 1$ given the observed value of H_n , and the prediction for $S_n = 1$ given the observed values of H_n and W_n . Naturally, the prediction for $S_n = 1$ given H_n , denoted $P_n(S_n = 1 | H_n)$, is computed using the Junction-tree FCVP algorithm with W, R, S as the labels and S as the target label, and the prediction for $S_n = 1$ given H_n and W_n , denoted $P_n(S_n = 1 | H_n, W_n)$, is computed using the Junction-tree FCVP with R, S as the labels and S as the target label.

Trials 3, 8, and 10, where some combination of H and W is observed for the first time, show the full conditionality of the predictor used: no informa-

Table 9.1. The FCVP as run on a data set randomly generated from the “wet grass” causal network: the first 17 trials, the trials in the range $n = 18, \dots, 100$ with $H_n = 1$, and the first trial from $n = 1000$ with $H_n = 1$ and $W_n = 1$; the latter is also given for three other data sets randomly generated from the same causal network (the superscript in square brackets indicates the initial state, if different from 0, of MATLAB’s pseudorandom numbers generator)

n	H_n	$P_n(S_n = 1 H_n)$	W_n	$P_n(S_n = 1 H_n, W_n)$	S_n
1	0	[0, 1]	0	[0, 1]	0
2	0	[0, 0.5]	0	[0, 0.5]	0
3	1	[0, 1]	0	[0, 1]	1
4	0	[0, 0.333]	0	[0, 0.333]	0
5	0	[0, 0.25]	0	[0, 0.25]	0
6	0	[0, 0.2]	0	[0, 0.2]	0
7	1	[0.5, 1]	0	[0.5, 1]	1
8	1	[0.667, 1]	1	[0, 1]	0
9	0	[0, 0.167]	0	[0, 0.167]	0
10	0	[0, 0.143]	1	[0, 1]	0
11	0	[0, 0.125]	0	[0, 0.125]	0
12	0	[0, 0.111]	0	[0, 0.111]	0
13	0	[0, 0.1]	0	[0, 0.1]	0
14	1	[0.5, 0.75]	0	[0.647, 1]	1
15	0	[0, 0.091]	0	[0, 0.091]	0
16	0	[0, 0.083]	1	[0, 0.56]	0
17	1	[0.6, 0.8]	1	[0.167, 0.583]	0
31	1	[0.5, 0.667]	1	[0.152, 0.434]	0
32	1	[0.429, 0.571]	1	[0.118, 0.339]	0
34	1	[0.375, 0.5]	1	[0.094, 0.275]	0
41	1	[0.333, 0.444]	0	[0.713, 1]	1
49	1	[0.4, 0.5]	1	[0.085, 0.238]	0
52	1	[0.364, 0.455]	1	[0.082, 0.213]	0
53	1	[0.333, 0.417]	1	[0.072, 0.188]	0
54	1	[0.308, 0.385]	1	[0.065, 0.169]	1
55	1	[0.357, 0.429]	1	[0.088, 0.189]	1
56	1	[0.4, 0.467]	0	[0.829, 1]	1
59	1	[0.438, 0.5]	1	[0.135, 0.231]	0
65	1	[0.412, 0.471]	1	[0.113, 0.202]	1
66	1	[0.444, 0.5]	1	[0.136, 0.222]	0
67	1	[0.421, 0.474]	1	[0.125, 0.205]	0
75	1	[0.4, 0.45]	1	[0.121, 0.194]	0
78	1	[0.381, 0.429]	1	[0.117, 0.185]	0
82	1	[0.364, 0.409]	1	[0.105, 0.169]	0
83	1	[0.348, 0.391]	1	[0.099, 0.159]	0
1001	1	[0.374, 0.378]	1	[0.189, 0.194]	0
1005 ^[1]	1	[0.282, 0.286]	1	[0.134, 0.139]	1
1000 ^[2]	1	[0.291, 0.295]	1	[0.141, 0.145]	0
1006 ^[3]	1	[0.311, 0.315]	1	[0.160, 0.164]	0

tion gathered for other combinations is used and the prediction is vacuous, $[0, 1]$. The predictions $P_n(S_n = 1 | H_n)$ for both subsequences $\{n : H_n = 0\}$ and $\{n : H_n = 1\}$ are identical to the predictions in the Bernoulli problem (discussed on p. 159). The predictions $P_n(S_n = 1 | H_n, W_n)$, however, are not so simple: see trials 14 and 17. The predictions for trials 31–83 clearly show the “explaining away” phenomenon: the conditional probability of $S_n = 1$ drops after Holmes learns that $W_n = 1$. The predictions for trials starting from 1000 (for four different randomly generated data sets) can be compared with the true conditional probabilities: Jensen (1996) computes that $S_n = 1$ with probability 0.339 given $H_n = 1$, and $S_n = 1$ with probability 0.161 given $H_n = 1$ and $W_n = 1$. Only one of the eight predictions covers the corresponding true value, but this is not surprising: our current situation (with a simple network and full conditioning) is not so different from the Bernoulli case (p. 159), and so one would expect the order of magnitude n^{-1} for the n th prediction’s diameter and $n^{-1/2}$ for the accuracy with which the true probabilities can be estimated. Despite this lack of coverage, the FCVP’s predictions, as we know, still agree perfectly with the observed frequencies.

9.6 Proofs and further information

Proof of Theorem 9.1

The proof is based on the usual device of reversing the direction of time (cf. §8.7) and the fact that the Venn predictor’s multiprobability predictions contain the true probabilities if viewed backwards.

We will prove the following result, which will easily yield Theorem 9.1 (p. 224).

Proposition 9.7. *Let F be any Venn predictor. For any nonnegative reversible game supermartingale G with $G(\square) = 1$ and any probability distribution P on \mathbf{Z}^N that agrees with the given OCM $(\Sigma, \square, \mathbf{Z}, F, B)$, there exists a nonnegative random variable ξ on \mathbf{Z}^N such that $\int \xi dP \leq 1$ and, for any data sequence $(x_1, y_1, \dots, x_N, y_N) \in \mathbf{Z}^N$,*

$$G(P_1, y_1, \dots, P_N, y_N) \leq \xi(x_1, y_1, \dots, x_N, y_N), \quad (9.11)$$

where P_n , $n = 1, \dots, N$, is the prediction generated by F from x_1, y_1, \dots, x_n .

Proof. Since G is reversible, (9.11) is equivalent to

$$G(P_N, y_N, \dots, P_1, y_1) \leq \xi(x_1, y_1, \dots, x_N, y_N), \quad (9.12)$$

where, as before,

$$P_n := F(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$$

is the Venn predictor's output. We will prove the stronger statement

$$G(p_N, y_N, \dots, p_1, y_1) \leq \xi(x_1, y_1, \dots, x_N, y_N), \quad (9.13)$$

where

$$p_n = p_n(t_{n-1}(x_1, y_1, \dots, x_{n-1}, y_{n-1}), (x_n, y_n)), \quad n = N, \dots, 1,$$

is the distribution of the labels y^* in the pairs $(\sigma_{n-1}^*, (x^*, y^*)) \in \Sigma_{n-1} \times \mathbf{Z}$ generated by $B_n(t_n(x_1, y_1, \dots, x_n, y_n))$ conditional on

$$(\sigma_{n-1}^*, (x^*, y^*)) \in A_n(t_{n-1}(x_1, y_1, \dots, x_{n-1}, y_{n-1}), (x_n, y_n));$$

in symbols,

$$p_n\{y\} := \frac{B_n(\{(\sigma_{n-1}^*, (x^*, y^*)) \in A_n(\sigma_{n-1}, (x_n, y_n)) : y^* = y\} | \sigma_n)}{B_n(A_n(\sigma_{n-1}, (x_n, y_n)) | \sigma_n)},$$

where

$$\begin{aligned} \sigma_{n-1} &:= t_{n-1}(x_1, y_1, \dots, x_{n-1}, y_{n-1}), \\ \sigma_n &:= t_n(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, y_n) \end{aligned}$$

and (A_n) is the partition that determines the Venn predictor. Inequality (9.13) implies (9.12) because, by the definition of a Venn predictor, $p_n \in P_n$. To complete the proof we will show that

$$\eta_n(x_N, y_N, \dots, x_1, y_1) := G(p_N, y_N, \dots, p_{N-n+1}, y_{N-n+1})$$

(with $\eta_0(x_N, y_N, \dots, x_1, y_1)$ understood as $G(\square) = 1$) is a P -supermartingale under a suitable choice of the filtration (\mathcal{F}_n) . (Our current probability space is $\Omega := \mathbf{Z}^N$, and so the $\eta_n(x_N, y_N, \dots, x_1, y_1)$ is a special case of our generic notation $\eta_n(\omega)$; see §A.6.)

Each σ -algebra \mathcal{F}_n , $n = 0, 1, \dots, N$, on the space of elementary events

$$\omega = (x_N, y_N, \dots, x_1, y_1)$$

is generated by the random elements $x_{N-n+1}, y_{N-n+1}, \dots, x_N, y_N$, the summary $t_{N-n}(x_1, y_1, \dots, x_{N-n}, y_{N-n})$, and the partition A_{N-n} (notice that η_n is measurable w.r. to \mathcal{F}_n). To show that η is a P -supermartingale, we need to establish $\mathbb{E}(\eta_n | \mathcal{F}_{n-1}) \leq \eta_{n-1}$, i.e.,

$$\begin{aligned} \mathbb{E}\Big(&G(p_N, y_N, \dots, p_{N-n+1}, y_{N-n+1}) \\ &\quad | x_{N-n+2}, y_{N-n+2}, \dots, x_N, y_N, \\ &\quad t_{N-n+1}(x_1, y_1, \dots, x_{N-n+1}, y_{N-n+1}), A_{N-n+1} \Big) \\ &\leq G(p_N, y_N, \dots, p_{N-n+2}, y_{N-n+2}) \end{aligned}$$

(the expectation being w.r. to P). Since P agrees with M , this is equivalent to

$$\begin{aligned} \mathbb{E}_B\left(G(p_N, y_N, \dots, p_{N-n+1}, y_{N-n+1}) \mid A_{N-n+1}\right) \\ \leq G(p_N, y_N, \dots, p_{N-n+2}, y_{N-n+2}), \end{aligned}$$

where $p_N, y_N, \dots, p_{N-n+2}, y_{N-n+2}$ are deterministic, and the expectation symbol \mathbb{E}_B refers to $(\sigma_{N-n}, y_{N-n+1})$ being drawn from $B_{N-n+1}(\sigma_{N-n+1})$ for a given $\sigma_{N-n+1} \in \Sigma_{N-n+1}$ and p_{N-n+1} (an A_{N-n+1} -measurable probability distribution on \mathbf{Y}) calculated from $(\sigma_{N-n}, y_{N-n+1})$ as the distribution of the labels in $A_{N-n+1}(\sigma_{N-n}, y_{N-n+1})$ under $B_{N-n+1}(\sigma_{N-n+1})$. By the definition of p_{N-n+1} , the last equality follows from

$$\begin{aligned} \int G(p_N, y_N, \dots, p_{N-n+1}, y_{N-n+1}) p_{N-n+1}(dy_{N-n+1}) \\ \leq G(p_N, y_N, \dots, p_{N-n+2}, y_{N-n+2}), \end{aligned}$$

which in turn follows from the definition of a game supermartingale (p. 150). Since η is a supermartingale, we can set

$$\xi(x_1, y_1, \dots, x_N, y_N) := \eta_N(x_N, y_N, \dots, x_1, y_1). \quad \square$$

To deduce the theorem from the proposition, set S_n in the definition of N -validity to $\mathbb{E}_P(\xi' | z_1, \dots, z_n)$, the conditional expectation of $\xi' := \xi + 1 - \int \xi dP$ given z_1, \dots, z_n under the probability distribution P .

Maximum-likelihood estimation in junction-tree models

This subsection will justify the reference to maximum-likelihood estimation in §9.4 (Lemma 9.5 on p. 230). We consider a junction-tree model (V, \mathcal{E}, Ξ) ; as we know (Corollary 9.6), the data-generating distribution can be assumed to satisfy the decomposition property (9.2) (p. 226).

Lemma 9.8. *For any edge s in the junction tree, the variables in the two maximal disjoint subsets of $V \setminus C_s$ divided by C_s are conditionally independent given C_s .*

(The two maximal disjoint subsets can be formally defined as $\cup_{u \in U_1} C_u \setminus C_s$ and $\cup_{u \in U_2} C_u \setminus C_s$, where U_1 and U_2 are the two maximal connected sets of vertices in the junction tree (U, S) with the edge s removed.)

Proof. This follows immediately from the definition of conditional independence: if we fix the values of the variables in C_s , (9.2) becomes the product of an expression involving variables of one subset and an expression involving variables of the other subset. \square

Lemma 9.9. *In the case of junction trees, (9.2) can be represented as*

$$Q\{a\} = \frac{\prod_{u \in U} Q\{z : z(v) = a(v), \forall v \in C_u\}}{\prod_{s \in S} Q\{z : z(v) = a(v), \forall v \in C_s\}}, \quad (9.14)$$

where the uncertainty $\frac{0}{0}$ is resolved to 0.

Proof. The proof is by induction on the size of the junction tree. Replace the left-hand side of (9.14) by $Q\{z : z(v) = a(v), \forall v \in V\}$; we no longer assume that $V := \cup_{u \in U} C_u$ includes all the variables. If the junction tree contains only one vertex (and no edges, which means that the denominator of (9.14) is 1), (9.14) is obvious.

Now suppose that (9.14) holds for a junction tree (U, S) ; let us prove that it also holds for the junction tree (U^*, S^*) obtained from (U, S) by adding another edge s and vertex w ; let $s = \{u, w\}$, where u is a vertex in U . Let $V := \cup_{u \in U} C_u$ and $V^* := \cup_{u \in U^*} C_u$. We have (with the second equality following from the previous lemma):

$$\begin{aligned} & Q\{z : z(v) = a(v), \forall v \in V^*\} \\ &= Q\{z : z(v) = a(v), \forall v \in V\} \\ &\quad \times Q\left(\{z : z(v) = a(v), \forall v \in V^* \setminus V\} \mid \{z : z(v) = a(v), \forall v \in V\}\right) \\ &= Q\{z : z(v) = a(v), \forall v \in V\} \\ &\quad \times Q\left(\{z(v) = a(v), \forall v \in C_w \setminus C_u\} \mid \{z(v) = a(v), \forall v \in C_w \cap C_u\}\right) \\ &= Q\{z : z(v) = a(v), \forall v \in V\} \\ &\quad \times \frac{Q\{z : z(v) = a(v), \forall v \in C_w\}}{Q\{z : z(v) = a(v), \forall v \in C_s\}}. \end{aligned}$$

This completes the proof. \square

Lemma 9.10. *Given a data sequence z_1, \dots, z_n , the maximum-likelihood model Q of the form (9.2) satisfies, for all configurations a on V ,*

$$Q\{a\} = \frac{\sigma_{\square}(a)}{n} \prod_{u \in U \setminus \{\square\}} \frac{\sigma_u(a)}{\sigma_{u'}(a)}, \quad (9.15)$$

where σ is the summary of z_1, \dots, z_n (cf. (9.6) on p. 230) and $\frac{0}{0} := 0$.

Proof. It is convenient to rewrite (9.14), analogously to (9.6), as

$$\begin{aligned} Q\{a\} &= Q\{z : z(v) = a(v), \forall v \in C_{\square}\} \\ &\times \prod_{u \in U \setminus \{\square\}} Q\left(\{z : z(v) = a(v), \forall v \in C_u\} \mid \{z : z(v) = a(v), \forall v \in C_{u'}\}\right). \end{aligned} \quad (9.16)$$

The proof of the lemma is again by induction. For a one-vertex junction tree the statement of the lemma follows from the fact that the maximum-likelihood multinomial probabilities coincide with the empirical frequencies.

Suppose (9.15) holds for a junction tree (U, S) . For a tree (U^*, S^*) obtained from (U, S) by adding a vertex u and an edge u' , we want to maximize

$$\begin{aligned} \prod_{i=1}^n Q\{z : z(v) = z_i(v), \forall v \in V^*\} &= \prod_{i=1}^n Q\{z : z(v) = z_i(v), \forall v \in V\} \\ &\times \prod_{i=1}^n Q(\{z : z(v) = z_i(v), \forall v \in C_u\} \mid \{z : z(v) = z_i(v), \forall v \in C_{u'}\}) \end{aligned}$$

(cf. (9.16)). It suffices to notice that we can maximize the two factors in the right-hand side independently and remember that, for a fixed configuration b on $C_{u'}$, the maximum-likelihood multinomial probabilities

$$Q(\{z : z(v) = a(v), \forall v \in C_u\} \mid \{z : z(v) = b(v), \forall v \in C_{u'}\}) ,$$

a ranging over the configurations on C_u that agree with b , are given by the empirical frequencies $\sigma_u(a)/\sigma_{u'}(b)$. \square

9.7 Bibliographical remarks

In our exposition of the hypergraphical model we mainly follow Vovk 2004. For further information about hypergraphs, see Lauritzen 1996 (§2.2).

Additive models

The hypergraphical model, as well as two of the models considered in the previous chapter (exchangeability model and Gaussian model) are additive in the following abstract sense: Σ is an Abelian semigroup and the statistics t_n satisfy

$$t_n(z_1, \dots, z_n) = t_1(z_1) + \dots + t_1(z_n) ,$$

where “+” is the semigroup operation. The theory of such repetitive structures (with uniform conditional distributions $P_n(\cdot \mid \sigma_n)$) is especially rich and is treated in Chap. III of Lauritzen 1988.

Perspectives and contrasts

This book has emphasized transductive methods for machine learning. In this concluding chapter, we step back to survey the historical and philosophical context of these methods, contrasting them with inductive and Bayesian methods.

We begin, in §10.1, by discussing inductive methods. We review the history of inductive learning under randomness, from Jacob Bernoulli's eighteenth-century law of large numbers to recent developments in statistical learning theory. All of this work has been done in the off-line framework, but at the end of the section we also discuss how the martingale approach to hypotheses testing can be used for on-line inductive learning.

In §10.2 we turn to transductive methods. We again take a historical perspective. Although the word “transduction” came into the lexicon of machine learning only recently (as we noted in Chap. 1, it was introduced by Vapnik), there are many earlier instances of transduction – many instances where old examples have been used in a relatively direct way to predict new examples. In some of these cases we can find instances of conformal prediction. Of particular interest is the prediction interval for a new observation based on Student’s t -distribution that Ronald A. Fisher published in 1935. Our historical review begins with this work, continues with later work on tolerance intervals, and then moves on to the Vapnik–Chervonenkis approach to transduction, first set forth in their 1974 monograph, and to subsequent work in statistical learning theory.

In general, an inductive method produces a prediction rule that can be applied to many new examples, and it aims for a high probability that the rule will predict with high accuracy. These goals are attractive but sometimes difficult or impossible to achieve. Transductive methods aim for less. Instead of trying to control two parameters – the desired accuracy and the probability of finding a rule with that accuracy – they try to control only the overall frequency of accurate predictions. This means that even when they succeed, they can be criticized for achieving less than an inductive method might achieve. But as we explain in the last subsection of §10.2, the cogency of this criticism

is doubtful in the on-line setting, where we use each rule only once. In the off-line setting, where we find a rule from old examples and apply it to many new examples, the repeated use of the rule gives empirical meaning to talk about its probability of predicting accurately. But in the on-line setting, where a rule for prediction, to the extent that it is even explicitly formulated, is used only once before being replaced by an improved rule, it may be meaningless to talk about the rule's probability of predicting accurately. It makes more sense to emphasize the overall frequency of accurate prediction.

We discuss Bayesian learning in §10.3. Viewed from our framework, which allows examples to be governed by any probability distribution Q on \mathbf{Z} , Bayesian learning requires us to make additional assumptions. First we assume that Q is one of a small family $(Q_\theta : \theta \in \Theta)$ of probability distributions on \mathbf{Z} (this is the *statistical model*), and then we adopt a probability distribution μ on Θ (this is the *prior distribution*) to express our probabilities for which Q_θ it is. Because of these additional assumptions, Bayesian learning lies outside the framework of this book. But it is currently very popular, and so it seems wise to explain briefly how it contrasts with our transductive methods. Our main conclusion will be unsurprising to anyone familiar with Bayesian learning: although Bayesian predictors are valid on their own terms, this validity depends on the additional assumptions being correct. When these assumptions are violated, Bayesian predictors may lack the kind of validity conformal predictors have. Perhaps more surprising is the fact that a prior distribution can be used to construct a conformal predictor, which gives predictions resembling the Bayesian predictions when the Bayesian assumptions are correct but, like any conformal predictor, is valid even if they are incorrect (cf. p. 102).

There is a slight shift in our notation in this chapter. In previous chapters, where we emphasized on-line transduction and therefore usually considered only one new example, we usually wrote z_1, \dots, z_{n-1} for the old examples and z_n for the new example. Now that we are also interested in the case where the same rule may be applied to many new examples, we will often write z_1, \dots, z_l for the old examples and z_{l+1}, \dots, z_{l+k} for the new examples.

10.1 Inductive learning

Induction means using old examples to formulate a rule for prediction that is then applied to new examples. When we are learning under randomness, the quality of the resulting predictions is affected by the randomness of both classes of examples. Because of the randomness of the new examples, there may be no prediction rule that performs perfectly, and because of the randomness of the old examples, we cannot even expect to find the prediction rule that performs best. As discussed in Chap. 1, a precise mathematical statement of what can be achieved typically involves two positive numbers, often denoted by ϵ and δ ; ϵ is a level of imperfection we are willing to tolerate in a

prediction rule, and δ bounds the probability we will fail to attain even this level. A typical theoretical result says that a certain method produces, with probability at least $1 - \delta$, a rule whose probability of error is at most ϵ .

The oldest result of this type is Jacob Bernoulli's theorem, which appeared in 1713, in his posthumous *Ars Conjectandi*. Abraham De Moivre improved on this result in 1733, obtaining what we now call the normal approximation to the binomial distribution, and Bernoulli's and De Moivre's results were subsequently generalized in many directions. In this section, we review the generalizations most pertinent to learning under unconstrained randomness, including Vapnik and Chervonenkis's uniform law of large numbers and more recent work on data-dependent bounds.

At the end of the section, we discuss a martingale approach to inductive prediction, inspired by the game-theoretic foundation for probability introduced in Shafer and Vovk 2001. For historical reasons, we postpone to §10.2 a discussion of tolerance regions, another important approach to inductive prediction.

To avoid the possibility of misunderstanding, we should mention that in this section, as in this whole book, we do not address the broad philosophical problem of induction that David Hume introduced in the eighteenth century. In the bulk of this book we are concerned only with learning under randomness, where induction means using randomly chosen examples to find a general rule for making predictions about future randomly chosen examples.

Jacob Bernoulli's learning problem

It would be misleading to say that Bernoulli worked with our concept of induction, because his work preceded the concept of a probability distribution. We might even think of his purpose as transductive: he wanted to predict whether an event would happen (or whether something is true or false) by looking at previous examples. But this project led him to formulate what we now recognize as the problem of estimating a probability p from previous randomly chosen examples, and so we can regard him as the founder of the inductive approach to learning under randomness.

Bernoulli likened his examples to pebbles drawn from an urn containing white and black pebbles in the ratio r to s , so that the probability of drawing a white pebble is $p = r/(r + s)$. The problem he considered was that of estimating the ratio r/s . But from the viewpoint of our framework, he was studying the problem of learning under randomness when there are no objects and the labels are binary. As usual, we take the labels to be 0 and 1 rather than white and black. Because there are no objects, we write $\mathbf{Z} = \{0, 1\}$. We write p for the probability of 1; this defines the probability distribution Q on \mathbf{Z} . We write z_1, \dots, z_l for the old examples.

Translated into these terms, Bernoulli's accomplishment was to show how to find, for any positive constants ϵ and δ and any $p \in [0, 1]$, a threshold $N(\epsilon, \delta, p)$ such that

$$Q^l \left\{ (z_1, \dots, z_l) : \left| \frac{1}{l} \sum_{i=1}^l z_i - p \right| \geq \epsilon \right\} \leq \delta \quad (10.1)$$

for any $l \geq N(\epsilon, \delta, p)$. This assumes only that the z_i are independent and equal to 1 with probability p , and it shows that, for large l ,

$$\hat{p}_l := \frac{1}{l} \sum_{i=1}^l z_i$$

is likely to be an accurate estimate of p . Results of this type, where counts or averages are shown to estimate probabilities or expected values, are now called laws of large numbers.

Bernoulli's threshold $N(\epsilon, \delta, p)$ was remarkably large. In the numerical example he gives at the end of *Ars Conjectandi*, where $p = 0.6$, $\epsilon = 0.02$, and $\delta = 1/1001$, it comes out to 25,550, a huge number in the context of the data sets available in the eighteenth century. As Stigler says (1986a, p. 77), it "was more than astronomical; for all practical purposes it was infinite". Bernoulli does not admit to any disappointment, but he ends the book soon after the number 25,550 appears.

Bernoulli's failure to produce a useable error bound has been repeated many times by subsequent authors. The error bounds produced by much of the leading theoretical work are too loose to be of practical value for available data sets, though authors seldom follow Bernoulli's example by providing numerical illustrations that make this shortcoming clear.

Bernoulli's bound can be improved substantially. The essential step in removing the slack was taken by Abraham De Moivre in 1733. De Moivre's theorem, now considered a special case of the central limit theorem, tells us how to calculate approximate probabilities for $z_1 + \dots + z_l$. In modern terminology, it says that this sum, scaled properly, has an approximately normal distribution. This allows us to approximate the probability in (10.1), not merely bound it. In 1925, Karl Pearson used De Moivre's theorem to show that the lowest valid value for $N(0.02, 1/1001, 0.6)$ – i.e., the lowest value of l for which

$$B_{0.6}^l \left\{ (z_1, \dots, z_l) : \left| \frac{1}{l} \sum_{i=1}^l z_i - 0.6 \right| \geq 0.02 \right\} \leq \frac{1}{1001}$$

holds – is approximately 6498. He also found, modifying Bernoulli's argument and using Stirling's formula, which was not known to Bernoulli, what he thought to be a rigorous valid value for $N(0.02, 1/1001, 0.6)$; Sirazhdinov later showed that the value given by Pearson has to be replaced by 6568 (see Prokhorov 1986, §8).

Even with the improvement brought by De Moivre's theorem, Bernoulli's approach has a drawback that might seem fatal from a rigorously logical point of view: the threshold $N(\epsilon, \delta, p)$ depends on p , which we are supposed

not to know. To estimate p we are told to take l to be at least $N(\epsilon, \delta, p)$, but to compute $N(\epsilon, \delta, p)$, we need to know p already. But this difficulty can be solved, and although the solution is awkward (see, e.g., Stuart et al. 1999, §§19.9–19.11), it usually produces a value for $N(\epsilon, \delta, p)$ not a great deal different from what we get after the fact for $N(\epsilon, \delta, k/l)$, where k is the observed number of 1s among z_1, \dots, z_l .

Although De Moivre's theorem eliminates the slack in Bernoulli's bounds, it produces values of $N(\epsilon, \delta, p)$ that are still embarrassingly large. Although we would always like to hope that the probability of error in a serious matter will be less than 1/1001, most decisions must be based on far fewer than 6568 observations. So statisticians learn to be content with $\delta = 0.01$ or $\delta = 0.05$.

Relatively weak levels of confidence, even when there is little slack in theoretical bounds, remain common in learning under unconstrained randomness, even though we often have two advantages over Bernoulli and De Moivre: much larger databases (at least in some applications) and information about the objects to help us predict the labels. As we have already mentioned, this suggests that it may be too ambitious to try to control both the level of accuracy ϵ and the probability of inaccuracy δ .

Another feature of Bernoulli's and De Moivre's results that we still see in the inductive approach to learning under unconstrained randomness is the logarithmic dependence of the required number of examples on the parameter δ . This is the ubiquitous $\ln \frac{1}{\delta}$ (see, e.g., (10.4) on p. 249 or Vapnik 1998). Bernoulli commented on this dependence, pointing out that the same number of additional examples is required every time we multiply the desired odds $(1 - \delta) : \delta$ by 10. When the desired odds are 1000 : 1 ($\delta = 1/1001$), the number of observations required by Bernoulli's bound is 25,550. When they are increased to 10,000 : 1 ($\delta = 1/10001$), this increases by 5708, to 31,258. When they are increased to 100,000 : 1 ($\delta = 1/100001$), it increases again by 5708, to 36,966. When we take the slack out of Bernoulli's bounds, the dependence is still approximately logarithmic: according to Sirazhdinov, the 6568 observations needed for odds 1000 : 1 goes up by about 2570 every time we multiply these odds by 10 – to 9142 for odds 10,000 : 1 and to 11,709 for odds 100,000 : 1.

As a modern example of the logarithmic dependence on δ , we can cite Hoeffding's inequality (see §A.7). In the case considered by Bernoulli, it says that

$$\mathbf{B}_p^l \left\{ (z_1, \dots, z_l) : \left| \frac{1}{l} \sum_{i=1}^l z_i - p \right| \geq \epsilon \right\} \leq 2 \exp(-2\epsilon^2 l) .$$

If we invert this inequality to solve Bernoulli's problem, we obtain

$$N(\epsilon, \delta, p) = \frac{1}{2\epsilon^2} \ln \frac{2}{\delta} .$$

This has the happy feature that it does not depend on p . But it is looser than Sirazhdinov's bound: instead of 6568 for $N(0.02, 1/1001, 0.6)$, it gives 9503.

Statistical learning theory

What we now call statistical learning theory was launched by Vapnik and Chervonenkis over 30 years ago, first in a short note published in 1968 and then in a full article, with proofs, published in 1971. In this work, they presented a very general and natural inductive method and showed that its performance can be guaranteed, for sufficiently long data sequences, by a uniform law of large numbers.

Let \mathcal{A} be a family of measurable subsets of \mathbf{Z} (in general an arbitrary measurable space). We say that \mathcal{A} *shatters* a finite set $Z \subseteq \mathbf{Z}$ if for any $Z' \subseteq Z$ there exists $A \in \mathcal{A}$ such that $Z' = Z \cap A$. We write $\text{VC}(\mathcal{A})$ for the cardinality of the largest finite set \mathcal{A} shatters, and we call $\text{VC}(\mathcal{A})$ the *VC dimension* of \mathcal{A} . When \mathcal{A} shatters arbitrarily large sets, $\text{VC}(\mathcal{A}) := \infty$.

The key result in Vapnik and Chervonenkis's theory is the inequality

$$\begin{aligned} Q^l \left\{ (z_1, \dots, z_l) : \sup_{A \in \mathcal{A}} \left| \frac{|\{i = 1, \dots, l : z_i \in A\}|}{l} - Q(A) \right| > \epsilon \right\} \\ < 4 \exp \left(\left(\frac{\text{VC}(\mathcal{A})(1 + \ln(2l/\text{VC}(\mathcal{A})))}{l} - (\epsilon - 1/l)^2 \right) l \right), \quad (10.2) \end{aligned}$$

which holds for every \mathcal{A} satisfying $0 < \text{VC}(\mathcal{A}) < l$ and for every probability distribution Q on \mathbf{Z} . (There are many variations on this inequality, none of them canonical. This particular version appears in Vapnik 1998, Theorem 4.4.)

Remark To see the potential of inequality (10.2), let us take \mathbf{Z} to be $\{0, 1\}$ and \mathcal{A} to be $\{\emptyset, A\}$, where A is the singleton $\{1\}$. In this case, $\text{VC}(\mathcal{A}) = 1$, we can write p for $Q(\{1\})$, and the inequality becomes

$$Q^l \left\{ (z_1, \dots, z_l) : \left| \frac{1}{l} \sum_{i=1}^l z_i - p \right| > \epsilon \right\} < 4 \exp (1 + \ln(2l) - (\epsilon - 1/l)^2 l)$$

(we added \emptyset to \mathcal{A} to ensure $\text{VC}(\mathcal{A}) \neq 0$). For fixed ϵ the right-hand side of this inequality can be made arbitrarily small by making l sufficiently large, and so we obtain another proof of Bernoulli's law of large numbers.

The inequality (10.2) is important not simply because it generalizes Bernoulli's theorem but because it does so uniformly in A . The right-hand side does not involve A , and so we can make the probability of a given deviation between $Q(A)$ and $|\{i = 1, \dots, l : z_i \in A\}|/l$, the empirical frequency of A in the first l examples, small uniformly in A by taking l large enough. This is the Vapnik–Chervonenkis *uniform law of large numbers*: if $\text{VC}(\mathcal{A}) < \infty$, then the empirical frequencies of the sets $A \in \mathcal{A}$ converge to their probabilities $Q(A)$ in probability uniformly.

It is easy to see that this uniform law of large numbers allows us to derive guarantees for inductive classification. Recall that in our framework with $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, classification is the case where $|\mathbf{Y}| < \infty$. We extend the concept of

VC dimension to a family \mathcal{F} of measurable functions $f : \mathbf{X} \rightarrow \mathbf{Y}$ by taking $\text{VC}(\mathcal{F})$ to be the VC dimension of the family of their graphs

$$\{(x, y) \in \mathbf{X} \times \mathbf{Y} : y = f(x)\}.$$

Choose such a family \mathcal{F} , together with parameters $\epsilon > 0$ and $\delta > 0$, and suppose that for at least one function f in \mathcal{F} , $f(x_i)$ predicts y_i correctly for most i . If $0 < \text{VC}(\mathcal{F}) < \infty$, then (10.2) suggests the following strategy: choose l such that the right-hand side, with $\text{VC}(\mathcal{A})$ replaced by $\text{VC}(\mathcal{F})$, does not exceed δ , and choose a function \hat{f} in \mathcal{F} that minimizes the empirical error

$$E(f) := \frac{|\{i = 1, \dots, l : y_i \neq f(x_i)\}|}{l}.$$

(This choice of f is often referred to as ‘‘empirical risk minimization’’.) We then know that \hat{f} ’s probability of error as a prediction rule will not exceed $E(\hat{f}) + \epsilon$ unless an unlikely event – an event of probability at most δ – has happened. The inequality (10.2) further implies that this probability of error will not exceed

$$\inf_{f \in \mathcal{F}} Q\{(x, y) \in \mathbf{X} \times \mathbf{Y} : y \neq f(x)\} + 2\epsilon$$

unless an event of probability 2δ has happened.

There are several useful versions of (10.2), including extensions that are relevant to regression rather than classification. Moreover, it has been shown that the inequality is optimal in several senses. For example, a finite VC dimension is necessary and sufficient for the uniform convergence of frequencies to corresponding probabilities (Vapnik 1998, Theorem 4.5).

A number of practical prediction methods recommend using prediction rules having a small empirical error and belonging to a class of finite VC dimension (such as a subclass of neural networks; see also the description of structural risk minimization on p. 249). So the Vapnik–Chervonenkis uniform law of large numbers provides an asymptotic justification for methods actually used. The inequality (10.2) is far too loose, however, to tell us how confident we should be in the accuracy of the specific predictions these methods make. To see the problem, consider the sample size l required in (10.2) to make both probability bounds nontrivial – i.e., less than 1. For any result of this kind to have nontrivial implications, l must exceed the VC dimension. It is difficult to assess VC dimension precisely for the classes used in practice, but they are undoubtedly huge. One indication is the bound for the VC dimension of a sigmoid neural network obtained by Karpinski and Macintyre (1995, 1997; see also Anthony 2003). This bound is roughly $(WN)^2$, where N is the number of computational units and W is the number of independent parameters. For LeNet 1, the first and smallest of the neural networks designed by Yann LeCun’s group for recognizing hand-written digits of the type in the USPS data set, $N = 4635$ and $W = 2578$, and so the bound exceeds 10^{14} .

Karpinski and Macintyre's bound can likely be tightened, but it is a very long way to a practical result.

The fact that the Vapnik–Chervonenkis uniform law of large numbers and similar results are usually too loose to give guidance about the confidence that we should have in predictions from actual data sets is, of course, well known. See, for example, NeuroCOLT 2002.

Vapnik and Chervonenkis's theory was partially rediscovered by Leslie G. Valiant (1984), whose work helped create a large community of computer scientists who have enriched the theory with analyses of computational complexity. Their work came to be known as PAC theory, because they obtained estimates that were Probably Approximately Correct. In one respect, unfortunately, PAC theory was more restrictive than Vapnik and Chervonenkis's theory. Vapnik and Chervonenkis assumed only that at least one of the functions in \mathcal{F} did a good job of prediction, and in this sense \mathcal{F} was only a tool for them. In PAC theory, it was assumed that the observed examples are exactly labeled by a function from $\mathcal{F} \subseteq \mathbf{Y}^{\mathbf{X}}$ – i.e., that there exists $f \in \mathcal{F}$ with $y_i = f(x_i)$ for all i , and this makes \mathcal{F} part of the statistical model for Reality. Now that the greater generality of Vapnik and Chervonenkis's viewpoint is clearly understood, the contributions of the PAC theorists are considered contributions to Vapnik and Chervonenkis's statistical learning theory.

Remark It is interesting that the Glivenko–Cantelli theorem, which some probabilists consider the fundamental result of mathematical statistics, is a special case of Vapnik and Chervonenkis's result, (10.2). The Glivenko–Cantelli theorem says that the empirical distribution function

$$F_l(t) := \frac{|\{i = 1, \dots, l : z_i \leq t\}|}{l}$$

for a random variable ζ (z_1, z_2, \dots are independent realizations of ζ) converges almost surely to ζ 's distribution function F . To derive this result from (10.2), we take \mathbf{Z} to be the real line and \mathcal{A} to be the family of all sets of the form $(-\infty, t]$, $t \in \mathbb{R}$, so that $\text{VC}(\mathcal{A}) = 1$ and thus

$$\begin{aligned} Q^l \left\{ (z_1, \dots, z_l) : \sup_{t \in \mathbb{R}} |F_l(t) - F(t)| > \epsilon \right\} \\ < 4 \exp \left(1 + \ln(2l) - (\epsilon - 1/l)^2 l \right). \quad (10.3) \end{aligned}$$

This inequality means that the empirical distribution function converges to the distribution function uniformly in probability, and the convergence almost surely required by Glivenko–Cantelli follows by the Borel–Cantelli lemma. The right-hand side of (10.3) is not too different from the $2 \exp(-2\epsilon^2 l)$ obtained specifically for this special case by Dvoretzky et al. (1956) and Massart (1990). For further details, see Devroye et al. 1996 (§12.8).

If we are interested in the case where the “hypothesis space” \mathcal{F} is of infinite VC dimension, it will often be possible to represent \mathcal{F} as the union of a nested

sequence of function classes $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subset \mathcal{F}$ of finite VC dimension. For example, if \mathcal{F} is the set of functions computable by neural networks, \mathcal{F}_k may be the set of functions computable by neural networks with no more than k neurons. In this situation, empirical risk minimization should be replaced by “structural risk minimization”; for details, see Vapnik 1998 (Chap. 6). We will briefly describe the main idea of structural risk minimization in the case of transduction after the statement of Proposition 10.1 (p. 259).

The quest for data-dependent bounds

The excessive looseness of bounds such as (10.2) is often attributed to their nonconditionality – i.e., the fact that they do not depend on the particular training set z_1, \dots, z_l at hand. We might hope to identify some training sets that are particularly informative, in the sense that they produce prediction rules with much better accuracy, and we might hope for a distribution Q that produces such favorable training sets with high probability.

One popular and particularly elegant way of obtaining data-dependent bounds, which comes close to being useful, is Littlestone and Warmuth’s sample compression approach (Littlestone and Warmuth 1986, Floyd and Warmuth 1995). Their theorem (Theorem 4.25 and its corollary Theorem 6.8 in Cristianini and Shawe-Taylor 2000) tells us, as a special case, that the probability of error on a new example by a support vector machine making a prediction with d support vectors is bounded with probability $1 - \delta$ by

$$\epsilon := \frac{1}{l-d} \left(d \left(1 + \ln \frac{l}{d} \right) + \ln \frac{l}{\delta} \right), \quad (10.4)$$

where l is the size of the training set. The full theorem is applicable to any “sample compression scheme”; support vector machines are only one instance, albeit a powerful one, of such a scheme. A similar result holds for regression problems (Cristianini and Shawe-Taylor 2000, Theorems 4.26, 4.28, and 4.30).

Though it is one of the tightest data-dependent bounds, Littlestone and Warmuth’s bound still falls short of being useful. To see this, we observe that for each of the ten classifiers in the problem of identifying digits in the USPS data set, (10.4) is approximately

$$\frac{1}{7291 - 274} 274 \left(1 + \ln \frac{7291}{274} \right) \approx 0.17$$

even when we ignore the term $\ln \frac{l}{\delta}$. (There are 7291 training examples, and the average for the 10 classifiers of the number of support vectors for the polynomial kernel of degree 3, the degree that gives the best predictive performance, is 274; see Table 12.2 in Vapnik 1998.) Thus the bound on the total probability of a mistake by one or more of the ten classifiers is 1.7, not a useful bound for a probability. There are more sophisticated schemes for multi-label classification using a binary classifier, but they involve separating unnatural

classes, such as odd and even digits, and so would lead to large numbers of support vectors.

Another popular way of obtaining data-dependent bounds is the much more recent PAC-Bayesian approach (McAllester 1998, McAllester 1999a, Seeger 2003, Langford and Shawe-Taylor 2003). For a review of other approaches, see Herbrich and Williamson 2002.

A whole different tack, which is often advocated but seems to offer little real promise, is to derive bounds that depend on the probability distribution Q , in the hope that the relevant aspects of Q might be estimated from the training set well enough to make the bounds usable. One example is provided by the inequality

$$\begin{aligned} Q^l \left\{ (z_1, \dots, z_l) : \sup_{A \in \mathcal{A}} \left| \frac{|\{i = 1, \dots, l : z_i \in A\}|}{l} - Q(A) \right| > \epsilon \right\} \\ < 4 \exp \left(\left(\frac{H_{\text{ann}}(2l)}{l} - (\epsilon - 1/l)^2 \right) l \right). \quad (10.5) \end{aligned}$$

This is a strengthening of (10.2), but the function H_{ann} , called the “annealed entropy”, depends on Q and as well as \mathcal{A} . Another example is the well-known upper bound

$$\frac{\mathbb{E} \mathcal{K}_{l+1}}{l+1} \quad (10.6)$$

on a support vector machine’s error probability. Here $\mathbb{E} \mathcal{K}_{l+1}$ is the expected number of support vectors among examples z'_1, \dots, z'_{l+1} randomly generated from Q^{l+1} . (See Vapnik 1998, Theorem 10.5; Theorems 10.6 and 10.7 are also of this type.) These kinds of bounds leave us with the problem of estimating an aspect of Q and bounding the probable error of this estimate, which seems more daunting than the problem with which we began.

The hold-out estimate

Because it has no objects, Bernoulli’s problem is of little interest to statistical learning theory, which emphasizes the use of complicated and informative objects. But as we have already remarked in Chap. 1, Bernoulli’s theorem nevertheless has an important role in statistical learning theory, because it is used when we estimate an error rate from a hold-out sample.

We can use the USPS data set to illustrate how reasonable the results obtained from the hold-out estimate are. The figures given in Appendix B show that the hold-out estimate gives reasonable results for this data set (at least producing values much less than one): assuming that the first half (4649 examples) gives a prediction rule whose error rate on the second half is 4%, the generalization error is bounded above by, approximately,

$$4\% + z_{1\%} \sqrt{\frac{4\% \times 96\%}{4649}} \approx 4.9\%$$

with probability 99%. As we already mentioned, a major disadvantage of this bound is that it does not depend on the object whose label is being predicted, and taking into account the quality of the object will weaken the bound.

Thirty years ago, George Barnard deplored the underuse of the hold-out estimate (Stone 1974):

The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number of cases where it is actually used.

His words still ring true today. The accuracies of the methods in statistical learning theory we have been discussing can be assessed much better by looking at their performance on a hold-out sample than by using the loose inequalities that would guarantee their good performance on data sets immensely larger than those available.

The hold-out estimate encounters several difficulties in practice, however:

1. Usually we do not obtain as good a prediction rule using only the training set as we would have obtained using the entire data set. So far as the development of the prediction rule is concerned, the test set is wasted.
2. The hold-out estimate of prediction accuracy uses only the performance on the test set. So far as the evaluation of the prediction rule is concerned, the training set is wasted.
3. The hold-out estimate gives a single probability of error that applies to all new examples, regardless of how difficult they are. This single probability of error would apply, for example, to all three images in Fig. 1.2 on p. 4.

The last problem can be partly overcome if we find, from the training set, a reasonable division of all objects into a few disjoint classes (perhaps just two, such as “clear images” and “blurred images”) and estimate for each class a probability of error from its percentage of wrongly classified test objects. This approach makes our predictions more *conditional* on the information provided by the object. But because it decreases the size of both the training set and the test set for each prediction, it aggravates the first two problems.

These difficulties provide one motivation for the new methods developed in this book.

On-line inductive learning

The work on inductive learning that we have been reviewing has been in an off-line setting. It is possible, however, to take advantage of an on-line setting for inductive learning. In this subsection we describe one simple on-line inductive predictor, whose domain of applicability is narrow but which sometimes has

advantages over natural transductive procedures within its narrow domain. The construction is similar to that of §7.2.

In §8.8 we compared two styles of modeling under uncertainty: statistical modeling and on-line compression modeling. Statistical modeling is a suitable starting point in inductive learning and on-line compression modeling in transductive. In this section we assume that the true probability distribution lies in a given statistical model $(P_\theta : \theta \in \Theta)$, where Θ is a set in a finite-dimensional Euclidean space. This is a restrictive assumption: the true probability distribution is known except several parameters, whereas in this book we emphasized the high-dimensional case, such as learning under unrestricted randomness for rich object spaces \mathbf{X} . For simplicity, we will concentrate on the special case of binary Markov chains (considered earlier in §§7.3, 8.6, and 8.8), but it will be clear that our argument only depends on standard regularity assumptions (the main ones being that P_θ is continuous as a function of θ and that the true parameter value can be consistently estimated from the data).

A *martingale predictor* for a statistical model $(P_\theta : \theta \in \Theta)$ on \mathbf{Z}^∞ is a family $(S_n^\theta : \theta \in \Theta, n \in \mathbb{N})$ of random variables on \mathbf{Z}^∞ and a family Γ^δ , $\delta \in (0, 1)$, of confidence predictors such that:

- S_n^θ , $n = 1, 2, \dots$, is a nonnegative P_θ -supermartingale (as defined on p. 170) with $S_0^\theta = 1$, for all $\theta \in \Theta$;
- the prediction sets

$$\Gamma^{\epsilon, \delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) := (\Gamma^\delta)^\epsilon(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$$

satisfy

$$\Gamma^{\epsilon, \delta_1}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \subseteq \Gamma^{\epsilon, \delta_2}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$$

whenever $\delta_1 \geq \delta_2$.

(In our applications S^θ will be P_θ -martingales.) We will say that such a martingale predictor is *valid* if the conditional probability under P_θ that

$$y_n \notin \Gamma^{\epsilon, \delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \quad \& \quad S_n^\theta(x_1, y_1, x_2, y_2, \dots) < 1/\delta$$

given $(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n)$ never exceeds ϵ . Roughly speaking, the conditional probability of error is required to be at most ϵ under all probability distributions that are not discredited at level δ ; remember that S_n^θ depends only on the first n examples $x_1, y_1, \dots, x_n, y_n$.

Each valid martingale predictor satisfies the following version of the definition of conservative validity given in §2.1: for all ϵ and δ in $(0, 1)$ and all $\theta \in \Theta$, the sequence of random variables

$$\begin{cases} \text{err}_n^{\epsilon, \delta}(\Gamma) & \text{if } S_n^\theta < 1/\delta \\ 0 & \text{otherwise} \end{cases}$$

is dominated under P_θ in distribution (as defined on p. 21) by a sequence of independent Bernoulli random variables with parameter ϵ . The only difference from the definition given in §2.1 is that we now allow testing of the probability distributions in the statistical model $(P_\theta : \theta \in \Theta)$.

There are simple ways to construct valid martingale predictors for regular statistical models that can be hoped to have reasonable properties of efficiency. For simplicity we will assume that \mathbf{Z} is finite and that Θ is Borel and has a finite and positive Lebesgue measure; we will also assume that the relevant regular conditional probabilities exist and are fixed. A natural family of martingales S_θ is

$$S_n^\theta(z_1, z_2, \dots) := \frac{\int P_\theta(z_1, \dots, z_n) \mu(d\theta)}{P_\theta(z_1, \dots, z_n)},$$

where μ is the uniform probability distribution on Θ and $P_\theta(z_1, \dots, z_n)$ is the probability that the first n examples generated by P_θ will be z_1, \dots, z_n . For each x_1, y_1, \dots, x_n and $\theta \in \Theta$ sort $y \in \mathbf{Y}$ in the order of decreasing conditional probability under P_θ that $y_n = y$ given x_1, y_1, \dots, x_n ; let $y_{(1)}, \dots, y_{(|\mathbf{Y}|)}$ be the sorted sequence. Define

$$\Gamma^{\epsilon|\theta}(x_1, y_1, \dots, x_n) := \{y_{(1)}, \dots, y_{(k)}\},$$

where k is the smallest integer such that the conditional probability under P_θ of the event $y_n \in \{y_{(1)}, \dots, y_{(k)}\}$ given x_1, y_1, \dots, x_n is at least $1 - \epsilon$. Finally, set

$$\begin{aligned} \Gamma^{\epsilon,\delta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) &:= \bigcup_{\theta \in \Theta} \left(\Gamma^{\epsilon|\theta}(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n) \right. \\ &\quad \left. \cap \{y \in \mathbf{Y} : S_\theta(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, y) < 1/\delta\} \right). \end{aligned}$$

The martingale predictor composed of (S_n^θ) and $(\Gamma^{\epsilon,\delta})$ is always valid. Moreover, for regular statistical models the P_θ -martingales S_n^θ will eventually reject all θ outside a small neighborhood of the true parameter value, and we will also have asymptotic efficiency. In particular, in the case of the Markov model this martingale predictor does not suffer from the problem discussed on p. 220.

10.2 Transductive learning

Transductive prediction and model testing are two sides of the same coin – two of the many ways we can use a procedure that tells us whether a data set agrees with a statistical model.

Testing: If we entertain an *a priori* possible model M for the data z_1, \dots, z_n , then seeing this data and detecting strong disagreement between it and the model will make us abandon the model.

Prediction: If we strongly believe in the model M and have not yet seen the data, or at least not all of it, then detecting strong disagreement between z_1, \dots, z_n and M allows us to predict that z_1, \dots, z_n will not happen. We have been particularly interested in the case where each z_i consists of two components, x_i and y_i , and we have so far seen only z_1, \dots, z_{n-1}, x_n .

The topic of this book is prediction, but we start this historical survey with testing. As we already mentioned, the first test was devised by John Arbuthnott in the early eighteenth century, but Arbuthnott's model was too simple ($\mathbf{B}_{1/2}^n$ with $\mathbf{B}_{1/2}$ the uniform distribution on $\{0, 1\}$) to have interesting implications for prediction. So we start with Student's work and its development by Fisher.

Student and Fisher

In this subsection we take \mathbf{Z} to be \mathbb{R} and use the notation introduced in §8.5. We will assume that z_1, z_2, \dots are independent N_{μ, σ^2} random variables.

In Student (§III) William S. Gosset, writing as "Student", correctly guessed the probability distribution of the ratio

$$\frac{\bar{z}_n - \mu}{\hat{\sigma}_n};$$

this ratio does not depend on σ and so can be used for testing, e.g., the hypothesis $\mu = 0$. Fisher derived Student's distribution rigorously by September 1912 (see Pearson 1968), but a demonstration was not published until much later. From the point of view of the general theory it is more natural to consider the ratio

$$\sqrt{n} \frac{\bar{z}_n - \mu}{\hat{\sigma}_n},$$

and the distribution of this ratio is now known as Student's t -distribution (with $n - 1$ degrees of freedom).

Student's result is not general enough to lead to an interesting prediction procedure. A more general test was proposed in Fisher's 1925 paper, where, after rigorously deriving Student's result, he treated the problem of comparing the means of two independent samples from the same normal distribution. He showed that if $z_1, \dots, z_l, z_{l+1}, \dots, z_{l+k}$ are generated from N_{μ, σ^2}^{l+k} , then

$$\sqrt{\frac{l k (l + k - 2)}{l + k}} \frac{\bar{z}_{(1)} - \bar{z}_{(2)}}{\sqrt{S_{(1)} + S_{(2)}}}, \quad (10.7)$$

where

$$\begin{aligned} \bar{z}_{(1)} &:= \frac{1}{l} \sum_{i=1}^l z_i, \quad \bar{z}_{(2)} := \frac{1}{k} \sum_{i=l+1}^{l+k} z_i, \\ S_{(1)} &:= \sum_{i=1}^l (z_i - \bar{z}_{(1)})^2, \quad S_{(2)} := \sum_{i=l+1}^{l+k} (z_i - \bar{z}_{(2)})^2 \end{aligned}$$

are the means and sums of squared deviations from the means for the two samples, has Student's t -distribution with $l + k - 2$ degrees of freedom. This result, as Fisher pointed out in 1935, does have implications for prediction. Setting k to 1 in (10.7), he obtained a result that we mentioned earlier, in Chap. 8:

$$\sqrt{\frac{l}{l+1}} \frac{z_{l+1} - \bar{z}_l}{\hat{\sigma}_l} \quad (10.8)$$

has Student's t -distribution with $l - 1$ degrees of freedom. So z_{l+1} will belong to the interval (8.10) (p. 200, with $n = l + 1$) with probability $1 - \epsilon$.¹

Fisher stated his conclusion concerning z_{l+1} more starkly than we have just done. From (10.8), he concluded that after observing z_1, \dots, z_l , we can attribute a fully known probability distribution to z_{l+1} – the distribution of

$$\bar{z}_l + \xi \sqrt{\frac{l+1}{l}} \hat{\sigma}_l, \quad (10.9)$$

where \bar{z}_l and $\hat{\sigma}_l$ are now known constants and ξ has the t -distribution with $l - 1$ degrees of freedom. Fisher called this the *fiducial probability distribution* for z_{l+1} . In earlier articles, starting in 1930, he had similarly derived “fiducial distributions” for the parameters of various statistical models.

Fisher's fiducial argument was the topic of vigorous discussion in the 1930s and 1940s. He defended it until he died in 1962. (See §V.4 of his book *Statistical Methods and Scientific Inference*, first published in 1956.) But by the end of the 1930s, most mathematical statisticians had rejected it in favor of Jerzy Neyman's more restrained interpretation of estimation and prediction intervals, which recognizes that these intervals can have desired unconditional frequency properties but do not have all the properties we expect from probability intervals (see, e.g., Kolmogorov 1942, §5). The problem, which we discuss more fully at the end of this section, is that after observing z_1, \dots, z_l we should be interested in the conditional distribution of (10.8) given z_1, \dots, z_l , and this is an unknown normal distribution, not the known t -distribution. Fisher conceded the nonconditionality of the distribution, in his 1935 paper and subsequently, and he saw some force in the objection. In §V.3 of his 1956 book, e.g., he wrote:

For verification, the original prediction must be held firmly in view. This, of course, is a somewhat unnatural attitude for a worker whose main preoccupation is to improve his ideas. It is perhaps for this reason that some teachers assert that statements of fiducial probability cannot be tested by observations.

Neyman developed his theory of confidence intervals only for parameters, not for new observations. But Fisher's idea of basing a prediction interval on

¹Another author, George Baker, also published this result in 1935, but he was much less influential than Fisher.

the t -distribution gained some currency, especially after the appearance of his 1956 book. The idea often appears in books on linear regression and it also appears in the 1974 textbook by Cox and Hinkley.

Fisher's discussion concerned a single prediction, not a sequence of predictions, and he seems not to have realized that the confidence predictor (8.10) makes errors independently at different trials and that the chosen significance level ϵ therefore has a clear frequentist interpretation as the limiting frequency of errors. As we saw in §8.5, prediction intervals (8.10) cover the true z_n with the correct frequency ϵ in the long run in a natural learning protocol. More generally, (10.9) holds from $l = 2$ onwards in the sense that there exists a sequence ξ_l of independent random variables having the t -distribution with $l - 1$ degrees of freedom, $l = 2, 3, \dots$, such that

$$z_{l+1} = \bar{z}_l + \xi_l \sqrt{\frac{l+1}{l}} \hat{\sigma}_l ;$$

this has a plethora of frequentist implications. It is true, however, that the conditional distribution of ξ_l given z_1, \dots, z_l is normal rather than Student's.

Tolerance regions

Tolerance regions (or, more fully, statistical tolerance regions) were introduced in the 1941 paper by Wilks as a tool of induction. Wilks was motivated by Walter Shewhart's (1931) ideas about industrial mass production.

In this subsection we are only interested in the case where objects are absent and $\mathbf{Z} = \mathbf{Y}$ is a Euclidean space. For simplicity we will only consider tolerance regions under the randomness assumption with the additional assumption that the probability distribution generating the individual examples is absolutely continuous (i.e., has a density w.r. to the Lebesgue measure), although there have been a lot of work for other statistical models and for discontinuous distributions.

Let $\epsilon, \delta \in (0, 1)$. A measurable function

$$\Gamma : \mathbf{Z}^* \rightarrow 2^\mathbf{Z} \tag{10.10}$$

(cf. (2.6) on p. 19) is called an (ϵ, δ) -tolerance predictor² if

$$\inf_Q Q^l \left\{ (z_1, \dots, z_l) \in \mathbf{Z}^l : Q(\Gamma(z_1, \dots, z_l)) \geq 1 - \epsilon \right\} = 1 - \delta ,$$

where $l \in \mathbb{N}$ and Q ranges over the absolutely continuous probability distributions on \mathbf{Z} . If ϵ and δ are small, the output $\Gamma(z_1, \dots, z_l)$ of the tolerance

²More standard terms are, e.g., “ $1 - \delta$ tolerance region for a proportion $1 - \epsilon$ ” (Fraser 1957) or “ $(1 - \epsilon)$ -content tolerance region at confidence level $1 - \delta$ ” (Fraser and Guttman 1956), but we prefer a simpler name.

predictor can be used for predicting the future z_i , $i = l + 1, l + 2, \dots$: the prediction is that $z_i \in \Gamma(z_1, \dots, z_l)$.

Wilks (1941) constructed tolerance predictors in the one-dimensional case and Wald (1943) extended Wilks's procedure to the multi-dimensional case. Wald's construction was generalized by Tukey (1947) and then further extended by, among others, Fraser (1951, 1953) and Kemperman (1956).

There is a transductive variety of tolerance regions (see Fraser and Guttman 1956, Fraser 1957, Guttman 1970; now they are referred to as prediction regions or prediction sets), and the ideas for constructing inductive tolerance regions carry over easily to the transductive case.

Let $\epsilon \in (0, 1)$. A measurable function (10.10) is called an ϵ -tolerance predictor³ if

$$Q^{l+1} \{ (z_1, \dots, z_l, z_{l+1}) \in \mathbf{Z}^{l+1} : z_{l+1} \in \Gamma(z_1, \dots, z_l) \} = 1 - \epsilon, \quad (10.11)$$

where $l \in \mathbb{N}$ and Q ranges over the absolutely continuous probability distributions on \mathbf{Z} . We will only give a version of Tukey's (1947) construction of ϵ -tolerance predictors (although Tukey was interested in (ϵ, δ) -tolerance predictors). Tukey's predictor is essentially the conformal predictor determined by the following nonconformity measure.

For each $n = 1, 2, \dots$, fix a sequence of measurable functions $\phi_{n,k} : \mathbf{Z} \rightarrow \mathbb{R}$, $k = 1, \dots, n$, and a sequence of real numbers $\alpha_{n,1}, \dots, \alpha_{n,n}$. We will assume that the Lebesgue measure of $z \in \mathbf{Z}$ satisfying $\phi_{n,k}(z) = c$ is zero for all n , k , and $c \in \mathbb{R}$. For any sequence z_1, \dots, z_n of n examples, the corresponding nonconformity scores (2.15) (p. 25) are defined as follows. Assign the nonconformity score $\alpha_{n,1}$ to all z_i at which $\max \phi_{n,1}(z_i)$ is attained and discard these z_i . Then assign the nonconformity score $\alpha_{n,2}$ to all z_i at which $\max \phi_{n,2}(z_i)$ is attained and discard these z_i . Repeat this procedure, finally assigning the nonconformity score $\alpha_{n,n}$ to all z_i at which $\max \phi_{n,n}(z_i)$ is attained; if there are no z_i left at some stage, do nothing. (With probability one, at each stage $\max \phi_{n,k}$ will be attained at exactly one z_i .)

Tukey proved a general result showing, in particular, that at each significance level ϵ the conformal predictor just described satisfies (10.11), provided ϵ has the form $i/(l+1)$ for an integer i . Fraser (1951) noticed that we can allow $\phi_{n,k}$ to depend on the maxima reached by $\phi_{n,1}, \dots, \phi_{n,k-1}$ in the procedure for computing nonconformity scores. Kemperman (1956) further noticed we can allow dependence on the examples where the maxima were reached, not only on the maxima themselves.

Takeuchi (whose idea is published in a rudimentary form in Takeuchi 1975 but was explained more fully in his seminars at Stanford University in the late 1970s) arrived at a general notion very similar to that of conformal predictor.

³Or “ $1 - \epsilon$ expectation tolerance region” in a more standard terminology (Fraser 1957).

Transduction in statistical learning theory

Whereas the vast majority of theoretical results in statistical learning theory are stated in the inductive framework, transduction was an important source of intuition from the very beginning of the theory. For example, the main technical tool of the early theory, the so-called “ghost sample” technique (the technique is described in, e.g., Vapnik 1998, §4.13, although without using this expression), is of transductive nature. The idea of transduction was described already in the first monograph (Vapnik and Chervonenkis 1974, Chap. VI, §§10–13) devoted to statistical learning theory. In this subsection we will be using a more recent exposition, given in Vapnik 1998 (Chap. 8); we will start from the simplest result, and discuss a more interesting case after Proposition 10.1.

As in the case of induction (see the description of inductivist statistical learning theory in §10.1), we start from a fixed family \mathcal{F} of measurable functions mapping the object space \mathbf{X} to the label space \mathbf{Y} ; we will assume $|\mathbf{Y}| < \infty$. Let Q be a probability distribution on \mathbf{Z} and l and k be two positive integer numbers (the interesting case is where $l \gg 1$ and $k \gg 1$). Suppose we are given l examples z_1, \dots, z_l and k unlabeled objects x_{l+1}, \dots, x_{l+k} ; our goal is to predict the latters’ labels y_{l+1}, \dots, y_{l+k} . We will say that z_1, \dots, z_l is the *training set* and x_{l+1}, \dots, x_{l+k} is the *working set*. Vapnik and Chervonenkis (see, e.g., Vapnik 1998, Theorem 8.2 and (8.15)) found a function

$$E = E(\epsilon, l, k, \{x_1, \dots, x_{l+k}\}, p) \quad (10.12)$$

such that, for all Q , l , k , and $\epsilon \in (0, 1)$,

$$\begin{aligned} Q^{l+k} \left\{ (z_1, \dots, z_{l+k}) : \frac{|\{i = l+1, \dots, l+k : f(x_i) \neq y_i\}|}{k} \leq \right. \\ E \left(\epsilon, l, k, \{x_1, \dots, x_{l+k}\}, \frac{|\{i = 1, \dots, l : f(x_i) \neq y_i\}|}{l} \right), \forall f \in \mathcal{F} \Big\} \\ \geq 1 - \epsilon. \end{aligned} \quad (10.13)$$

Equation (10.13) is a transductive analog of (10.2) (p. 246), and it is applied to the problem of prediction in a similar way. Let ϵ be a small positive constant and suppose that at least one function from \mathcal{F} provides a good prediction rule. If E in (10.12) is reasonably small for given ϵ and training and working sets, we can apply the following strategy for predicting y_{l+1}, \dots, y_{l+k} . Choose a function $f = \hat{f} \in \mathcal{F}$ with the smallest

$$E \left(\epsilon, l, k, \{x_1, \dots, x_{l+k}\}, \frac{|\{i = 1, \dots, l : f(x_i) \neq y_i\}|}{l} \right) \quad (10.14)$$

(which typically means choosing an f with the best performance

$$\frac{|\{i = 1, \dots, l : f(x_i) \neq y_i\}|}{l}$$

on the training set); (10.13) implies that, unless an unlikely (of probability at most ϵ) event has occurred, the frequency of errors

$$\frac{|\{i = l + 1, \dots, l + k : f(x_i) \neq y_i\}|}{k} \quad (10.15)$$

on the working set will be bounded by (10.14).

Kolmogorov's objection (mentioned on p. 255) to Fisher's fiducial probabilities is also applicable to the Vapnik–Chervonenkis approach to transduction: it appears that we should be interested in the *conditional* probability given z_1, \dots, z_l (and perhaps also x_{l+1}, \dots, x_{l+k}) that the bound (10.14) is satisfied. One important advantage of conditional probabilities, however, does carry over to the bound (10.14). Let ϵ be a small positive constant (the probability of error we are willing to tolerate), $z_1 = (x_1, y_1), z_2 = (x_2, y_2), \dots$ be the observed sequence of examples, and l_1, l_2, \dots be a strictly increasing sequence of positive integers. Consider the following scenario of repeated Vapnik–Chervonenkis transduction. First we are given examples z_1, \dots, z_{l_1} and are asked to predict the labels of new objects $x_{l_1+1}, \dots, x_{l_2}$. We know that the bound (10.14), where $l := l_1$ and $k := l_2 - l_1$, will hold with probability ϵ . Next we are told the true labels for $x_{l_1+1}, \dots, x_{l_2}$, so we now know the full examples z_1, \dots, z_{l_2} . We are asked to predict the labels of new objects $x_{l_2+1}, \dots, x_{l_3}$. We again know that the bound (10.14), where $l := l_2$ and $k := l_3 - l_2$, holds with probability ϵ . We are now told the true labels for $x_{l_2+1}, \dots, x_{l_3}$, etc. If ϵ were an upper bound on the conditional probability that the bound (10.14) holds, we could deduce from the martingale strong law of large numbers that the limiting (in the sense of \limsup) frequency with which the bound (10.14) is violated does not exceed ϵ . Vapnik and Chervonenkis's result (10.13) by itself does not prevent this limiting frequency (even in the sense of \liminf) from being 1. However, using our standard methods we can deduce the following proposition, which asserts much more, namely, the conservative validity of the procedure based on (10.14) in the on-line protocol. Let err_n^ϵ be 1 if the bound (10.14) holds for the training set z_1, \dots, z_{l_n} and the working set $x_{l_n+1}, \dots, x_{l_{n+1}}$, and let it be 0 otherwise.

Proposition 10.1. *Let $\epsilon \in (0, 1)$. Suppose (10.13) holds with Q^{l+k} replaced by any exchangeable probability distribution P on \mathbf{Z}^{l+k} , for all l and k . In the on-line transduction protocol described above, err_n^ϵ , $n = 1, 2, \dots$, are dominated in distribution by independent Bernoulli random variables with parameter ϵ ; in particular,*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{err}_i^\epsilon \leq \epsilon.$$

The condition of the proposition ((10.13) holding with Q^{l+k} replaced by any exchangeable probability distribution on \mathbf{Z}^{l+k}) is satisfied for the usual choices of the function E in (10.12); for details, see §10.4.

The simplest case of Vapnik–Chervonenkis transduction is not very interesting from the point of view of statistical learning theory: for example,

the chosen prediction rule \hat{f} does not usually depend on the working set. Our analysis, however, can be easily extended to more advanced results, such as (8.35) in Vapnik 1998. In the rest of this section we will briefly describe the procedure of “structural risk minimization” in the case of transduction.

Suppose for each choice of the training set x_1, \dots, x_l of objects and the working set x_{l+1}, \dots, x_{l+k} we have a representation of the function class \mathcal{F} as the union of an increasing sequence of function classes

$$\mathcal{F}_1^{\{x_1, \dots, x_{l+k}\}} \subseteq \mathcal{F}_2^{\{x_1, \dots, x_{l+k}\}} \subseteq \dots \subseteq \mathcal{F}$$

(depending only on the combined set, or more accurately bag, $\{x_1, \dots, x_{l+k}\}$). For example, the classes $\mathcal{F}_r^{\{x_1, \dots, x_{l+k}\}}$ for small r may consist of the functions that separate the combined set $\{x_1, \dots, x_{l+k}\}$ with a large margin (as in Vapnik 1998, §8.5, Vapnik and Chervonenkis 1974, §VI.10). Suppose we have a function

$$E = E(\epsilon, l, k, \{x_1, \dots, x_{l+k}\}, p, r)$$

such that, for all l and k , all exchangeable distributions P on \mathbf{Z}^{l+k} , and all $\epsilon \in (0, 1)$,

$$\begin{aligned} P\left\{ (z_1, \dots, z_{l+k}) : \frac{|\{i = l+1, \dots, l+k : f(x_i) \neq y_i\}|}{k} \leq \right. \\ \left. E\left(\epsilon, l, k, \{x_1, \dots, x_{l+k}\}, \frac{|\{i = 1, \dots, l : f(x_i) \neq y_i\}|}{l}, r\right), \right. \\ \left. \forall r, \forall f \in \mathcal{F}_r^{\{x_1, \dots, x_{l+k}\}} \right\} \geq 1 - \epsilon \quad (10.16) \end{aligned}$$

(for examples of such E , see Vapnik 1998, Theorem 8.4 and (8.35), and Vapnik and Chervonenkis 1974, Theorem 6.2).

Equation (10.16) is used for prediction in a more interesting way than (10.13). Let ϵ be a small positive constant. To predict y_{l+1}, \dots, y_{l+k} , choose r and a function $f = \hat{f} \in \mathcal{F}_r$ with the smallest

$$E\left(\epsilon, l, k, \{x_1, \dots, x_{l+k}\}, \frac{|\{i = 1, \dots, l : f(x_i) \neq y_i\}|}{l}, r\right); \quad (10.17)$$

(10.16) implies that, unless an unlikely (of probability at most ϵ) event has occurred, the frequency of errors (10.15) on the working set will be bounded by (10.17) (with f replaced by \hat{f}). It is easy to check that Proposition 10.1 can be stated and proved for this “structural risk minimization” framework.

PAC transduction

The main work on transduction in the PAC tradition was done by Haussler, Littlestone, and Warmuth (1994); it will be described in this subsection.

Following Haussler et al. (1994) we assume that the label space is binary, $\mathbf{Y} = \{0, 1\}$, and fix a class \mathcal{F} of functions of the type $\mathbf{X} \rightarrow \mathbf{Y}$ of finite VC dimension. For simplicity, we will also assume that the object space \mathbf{X} is finite and that the probability distribution $Q \in \mathbf{P}(\mathbf{Z})$ generating the individual examples satisfies $Q(\{x, y\}) > 0$ for all $(x, y) \in \mathbf{X} \times \mathbf{Y}$.

In the model considered by Haussler et al. (1994), and adopted in this subsection, the examples are generated by the power probability distribution Q^∞ with the probability distribution Q assumed compatible with \mathcal{F} (i.e., Q is such that, for some $f \in \mathcal{F}$, $f(x) = y$ for all $(x, y) \in \mathbf{Z}$). Because of the restriction to probability distributions compatible with \mathcal{F} , we will have to modify the definition of a conformal predictor. The \mathcal{F} -conformal predictor determined by a nonconformity measure (A_n) is the following confidence predictor: (2.17) (p. 26) is defined to be the set of all labels $y \in \mathbf{Y}$ such that

- the data sequence $(x_1, y_1, \dots, x_{n-1}, y_{n-1}, x_n, y)$ is *compatible* with \mathcal{F} , in the sense that there exists $f \in \mathcal{F}$ such that $f(x_i) = y_i$, $i = 1, \dots, n-1$, and $f(x_n) = y$,
- (2.18) holds, where the nonconformity scores α_i are defined by (2.19).

The smoothed version of this definition can be given and the analogues of Propositions 2.3 and 2.4 proved. Theorem 2.2 in Haussler et al. 1990 (an early version of Haussler et al. 1994 that explicitly states several key results) can be restated in the following way:

Proposition 10.2 (Haussler, Littlestone, and Warmuth). *Let \mathcal{F} be a class of $\{0, 1\}$ -valued functions on \mathbf{X} of finite VC dimension; consider the variable significance level $\epsilon_n := 2 \text{VC}(\mathcal{F})/n$. There exists an \mathcal{F} -conformal predictor Γ such that for any sequence $(x_1, y_1, x_2, y_2, \dots)$ of examples there are no multiple predictions:*

$$|\Gamma^{\epsilon_n}(x_1, y_1, \dots, x_n)| \leq 1, \quad n = 1, 2, \dots. \quad (10.18)$$

Of course, the statement of the proposition is of interest only for sequences of examples that are compatible with \mathcal{F} : there is $f \in \mathcal{F}$ such that $y_i = f(x_i)$, $i = 1, 2, \dots$. The proposition will be proved in §10.4.

The restatement given by Proposition 10.2 sheds new light on the performance of Haussler et al.'s procedure in repeated trials (see the discussion at the end of §2 in Haussler et al. 1994, in particular Corollary 2.2): the independence of the errors at different trials allows us to use the powerful machine of probability theory. We will only give two examples.

The on-line \mathcal{F} -conformal predictor Γ whose existence is asserted in Proposition 10.2 never makes multiple predictions at the significance level $2 \text{VC}(\mathcal{F})/n$, and so the only measure of its performance at this level is the number of errors

$$\text{Err}_n := \text{Err}_n^{2 \text{VC}(\mathcal{F})/n}(\Gamma)$$

it makes. The law of the iterated logarithm (p. 287) implies that,

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n - 2 \text{VC}(\mathcal{F}) \ln n}{2\sqrt{\text{VC}(\mathcal{F}) \ln n \ln \ln \ln n}} \leq 1 \quad \text{a.s.}$$

Poisson's theorem (see, e.g., Shiryaev 1996, Theorem III.3.4) shows that the distribution of the cumulative number of errors $\text{Err}_N - \text{Err}_{[N/2]}$ in the second half of the first N trials is asymptotically dominated, as $N \rightarrow \infty$, by the Poisson distribution with the parameter $(2 \ln 2) \text{VC}(\mathcal{F})$.

Why on-line transduction makes sense

The fact that traditional off-line learning theory is inductive is not purely accidental: for reasons that we will now discuss, transductive learning is relatively ill-suited to the off-line framework. Only in the on-line framework do the error probabilities guaranteed by the theory find their manifestation as frequencies.

Suppose we are interested in the quality of an algorithm that takes examples z_1, \dots, z_l and an object x , and outputs a prediction set $F(z_1, \dots, z_l, x)$ for x 's label y . In the case of conformal prediction, we might fix a significance level ϵ and set

$$F(z_1, \dots, z_l, x) := \Gamma^\epsilon(z_1, \dots, z_l, x)$$

for some conformal predictor Γ . In the case of statistical learning theory, we might find a prediction rule $f_{z_1, \dots, z_l} : \mathbf{X} \rightarrow \mathbf{Y}$ for each sequence z_1, \dots, z_l and set $F(z_1, \dots, z_l, x)$ to the one-element set $\{f_{z_1, \dots, z_l}(x)\}$.

From an inductive viewpoint, F is a reliable predictor if there are small positive constants ϵ and δ such that, for all probability distributions Q on \mathbf{Z} ,

$$Q^l \{(z_1, \dots, z_l) : Q \{(x, y) : y \notin F(z_1, \dots, z_l, x)\} \leq \epsilon\} \geq 1 - \delta. \quad (10.19)$$

From a transductive viewpoint, F is reliable if there is a small positive constant ϵ such that, for all probability distributions Q on \mathbf{Z} ,

$$Q^{l+1} \{(z_1, \dots, z_l, (x, y)) : y \notin F(z_1, \dots, z_l, x)\} \leq \epsilon. \quad (10.20)$$

In other words, in the inductive approach we are interested in the random variable

$$\xi = \xi(z_1, z_2, \dots) := Q \{(x, y) : y \notin F(z_1, \dots, z_l, x)\}$$

(we want it to be small with high probability), whereas in the transductive approach we are interested in the expected value $\mathbb{E}\xi$ of ξ over the random choice of the training set z_1, \dots, z_l . Is this averaging over different training sets appropriate? It is generally agreed that in the off-line framework knowing $\mathbb{E}\xi$ is only marginally useful; as Devroye et al. (1996, p. 2) put it, "this number would indicate the quality of an average data sequence, not *your* data sequence". Kolmogorov's objection against Fisher's fiducial probability (p. 255) is also a version of this general *inductivist objection*, as we call it.

We will discuss separately two aspects of the inductivist objection, one more practical and the other more philosophical. Both aspects are very convincing in the off-line framework and both weaken in the on-line framework.

The practical aspect of the inductivist objection is that the knowledge that, say, $\mathbb{E}\xi = 5\%$ does not mean that we will be wrong in 5% of cases applying $F(z_1, \dots, z_n, x)$ to many new examples (x, y) generated independently from Q . Indeed, it might happen that, e.g., $\xi = 0$ with probability 0.75 and $\xi = 20\%$ with probability 0.25. If our training set is one of the one-in-four unlucky ones, we will make errors in 20% of all cases. The average probability 5% reflects what might have happened at the stage when z_1, \dots, z_l was generated, but we are interested in our particular z_1, \dots, z_l : is it a lucky one?

The philosophical aspect of the inductivist objection is that we want to know the true error probability. The average probability $\mathbb{E}\xi$ ceases to be the true error probability as soon as the training set z_1, \dots, z_l is generated. We should condition properly on what we already know. In the off-line setting the philosophical aspect also has a practical side: feeding $F(z_1, \dots, z_n, x)$ with sufficiently many new examples (x, y) , we might hope that the truth will be eventually (at least in the limit) revealed to us, and any deviation from the truth will become visible and practically significant.

There is a chance, of course, that even a predictor satisfying (10.19) will mislead us and the actual probability of error will exceed ϵ . But the situation here appears under control: we know that we can be misled only with a low probability, δ .

Things change in the on-line setting, where we take $(z_1, \dots, z_l) := (z_1, \dots, z_{n-1})$ and $(x, y) := z_n$ for $n = 1, 2, \dots$. As the practical aspect is concerned, we know from Chap. 2 that, in the case of conformal prediction, the error probability ϵ in (10.20) translates with high probability into the frequency of errors less than or close to ϵ . (Moreover, as we know from Chap. 4, it is sufficient to have some, maybe quite limited, supply of fresh examples, not necessarily arriving at each trial.) On the other hand, (10.19) becomes difficult to interpret when it is applied at every trial.

The notion of a true probability distribution becomes much murkier in the on-line framework. Even if we do choose a prediction rule, it becomes updated very quickly, and we simply do not have time to discover the truth.

Let us illustrate this on the simplest situation, where x_n are absent and $\mathbf{Y} = \{0, 1\}$. Suppose p is chosen randomly from the uniform distribution on $[0, 1]$ and then z_1, z_2, \dots are generated independently from the Bernoulli distribution \mathbf{B}_p with parameter p . The mixture $B := \int \mathbf{B}_p^\infty dp$ is Laplace's rule of succession (mentioned earlier in §6.3 and going back to Laplace 1774), which can be described in terms of its conditional probabilities: the B -probability that $z_{l+1} = 1$ given z_1, \dots, z_l is $(k+1)/(l+2)$, where k is the number of 1s among z_1, \dots, z_l . What is the true probability that $z_{l+1} = 1$ given z_1, \dots, z_l ? Is it p or is it $(k+1)/(l+2)$? The answer seems to depend on whether we first generated p and then generated the data from \mathbf{B}_p^∞ or we generated the data directly from B . There is no way, however, to tell from the data

alone in which of these two ways they were generated. It might even happen that only part of p was generated: e.g., we first chose randomly an interval $[i/10, (i+1)/10]$, $i = 0, \dots, 9$, and then generated the data from the mixture $(1/10) \int_{i/10}^{(i+1)/10} \mathbf{B}_p^\infty dp$.

If we apply the Venn predictor, $[k/(l+1), (k+1)/(l+1)]$, to estimate the probability that $z_{l+1} = 1$ in the situation of the previous paragraph, it is true that there is a good chance that the interval $[k/(l+1), (k+1)/(l+1)]$ will not cover the probability p (the typical distance between this interval and p has the order of magnitude $l^{-1/2}$). This interval, however, always covers another probability, Laplace's $(k+1)/(l+2)$. What matters is whether our predictor is valid and efficient, not what it claims to tell us about unobservable aspects of the genesis of the data.

10.3 Bayesian learning

The Bayesian approach to learning was suggested independently by Thomas Bayes, in a paper published posthumously in 1763, and by Pierre Simon Laplace, in a memoir published in 1774 (Stigler 1986b). Since then it has always had its advocates, but it has become particularly popular in the last third of the twentieth century. The main idea of the Bayesian approach is to complement whatever model $(P_\theta : \theta \in \Theta)$ for the data we might have by a new component (the prior $\mu(d\theta)$ for the parameters) so that a complete probability distribution, $\int P_\theta \mu(d\theta)$, for the data is obtained. With this probability distribution, the problem of prediction can be solved automatically by computing predictive distributions for new examples; learning from experience is also done automatically by applying Bayes's theorem. Therefore, the Bayesian approach is formally outside the scope of this book, which is about learning under uncertainty and does not assume the knowledge of the probability distribution for the data. It is still informative, however, to look at how the Bayesian approach for different plausible probability distributions compares with our approach; this is what we will do in this section. When the chosen probability distribution is the real one generating the data (realistically, we can know this only for artificial data sets that we ourselves generate), the Bayesian method can be counted on to give valid results; we will see in this section how the validity of those results is affected when the chosen distribution is wrong. By the results of Chaps. 2–4 conformal predictors, even if motivated by Bayesian considerations (we will see that RRCMs can be interpreted this way), are automatically valid.

Bayesian ridge regression

We start by giving a standard Bayesian derivation of the ridge regression estimator; this derivation will also provide us with the full conditional distribution

for the new label y_{l+1} . As usual, we have training examples $(x_1, y_1), \dots, (x_l, y_l)$ and a new object x_{l+1} , and our goal is to predict y_{l+1} ; the object space is $\mathbf{X} = \mathbb{R}^p$. Let us assume that the objects x_1, x_2, \dots are fixed (deterministic) and the labels y_1, y_2, \dots are generated by the rule

$$y_i = w \cdot x_i + \xi_i \quad (10.21)$$

(already encountered on pp. 35 and 201), where w is distributed as $\mathbf{N}_{0, (\sigma^2/a)I_p}$, each ξ_i is distributed as \mathbf{N}_{0, σ^2} , and all these random elements are independent.

The posterior density of w is proportional to

$$\begin{aligned} & \exp\left(-\frac{a}{2\sigma^2}\|w\|^2\right) \prod_{i=1}^l \exp\left(-\frac{1}{2\sigma^2}(y_i - w \cdot x_i)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(a\|w\|^2 + \sum_{i=1}^l (y_i - w \cdot x_i)^2\right)\right). \end{aligned} \quad (10.22)$$

It attains its maximum at the w , which we will denote \hat{w} , that solves the optimization problem (2.25) (p. 29, with $n = l$), i.e., at the value given by the ridge regression procedure.

Using the notation X_l and Y_l introduced in Chap. 2 (see (2.26) and (2.27) on p. 30), we can rewrite the right-hand side of (10.22) as

$$\begin{aligned} & \exp\left(-\frac{1}{2\sigma^2}(w'(X'_l X_l + aI_p)w - 2Y'_l X_l w + Y'_l Y_l)\right) \\ & \propto \exp\left(-\frac{1}{2\sigma^2}(w - \hat{w})'(X'_l X_l + aI_p)(w - \hat{w})\right). \end{aligned} \quad (10.23)$$

This can be recognized as the multivariate normal distribution with mean \hat{w} and variance matrix $V := \sigma^2(X'_l X_l + aI_p)^{-1}$ (see, e.g., Shiryaev 1996, §II.13); we are primarily interested, however, not in the parameter w but in the next label

$$y_{l+1} = w \cdot x_{l+1} + \xi_{l+1}.$$

The conditional distribution of $w \cdot x_{l+1}$ (given the training examples) is

$$\begin{aligned} & \mathbf{N}(\hat{w} \cdot x_{l+1}, x'_{l+1} V x_{l+1}) \\ &= \mathbf{N}(x'_{l+1}(X'_l X_l + aI_p)^{-1} X'_l Y_l, \sigma^2 x'_{l+1}(X'_l X_l + aI_p)^{-1} x_{l+1}) \end{aligned}$$

(the expression for \hat{w} is from (2.29); we write $\mathbf{N}(\mu, \sigma^2)$ for $\mathbf{N}_{\mu, \sigma^2}$ if the expression for μ or σ^2 is complicated). The assumption of independence now gives the conditional distribution

$$\mathbf{N}(x'_{l+1}(X'_l X_l + aI_p)^{-1} X'_l Y_l, \sigma^2 x'_{l+1}(X'_l X_l + aI_p)^{-1} x_{l+1} + \sigma^2) \quad (10.24)$$

for y_{l+1} .

All our experiments will be performed for the Bayesian procedure given by (10.24), but for completeness we will also give a kernel version of this procedure. It is possible to rewrite (10.24) in the kernel form using the matrix equation (2.39) (p. 35), similarly to what we did in Chap. 2 to derive the kernel representation of RRGM. (This is done in Melluish et al. 2001a and Melluish 2005.) It is easier, however, to compute the predictive distribution for y_{l+1} directly. Namely, it is easy to see that the covariance between y_i and y_j ($i, j = 1, \dots, l+1$) under the model

$$y_i = w \cdot F(x_i) + \xi_i$$

(which is (10.21) in the feature space; we are using the same notation as in §2.3 for the mapping $F : \mathbf{X} \rightarrow \mathbf{H}$ to the feature space and for the kernel \mathcal{K}) is given by

$$\text{cov}(y_i, y_j) = \frac{\sigma^2}{a} \mathcal{K}(x_i, x_j) + \sigma^2 \mathbb{I}_{i=j}; \quad (10.25)$$

therefore, the theorem on normal correlation (see, e.g., Shiryaev 1996, Theorem II.13.2) gives the predictive distribution

$$\mathbf{N} \left(Y'_l (K_l + aI_l)^{-1} k_l, \sigma^2 + \frac{\sigma^2}{a} \mathcal{K}(x_{l+1}, x_{l+1}) - \frac{\sigma^2}{a} k'_l (K_l + aI_l)^{-1} k_l \right) \quad (10.26)$$

for y_{l+1} , where K_l and k_l are essentially the same matrix and vector as in (2.41) (p. 36): $(K_l)_{i,j} := \mathcal{K}(x_i, x_j)$, $i, j = 1, \dots, l$, $(k_l)_i := \mathcal{K}(x_{l+1}, x_i)$, $i = 1, \dots, l$, and \mathcal{K} is defined by (2.42).

It is clear that at each significance level ϵ the shortest prediction interval is the one symmetrical w.r. to the ridge regression prediction \hat{y}_{l+1} . Namely, (10.24) gives the prediction interval

$$[\hat{y}_{l+1} - z_{\epsilon/2} V_l, \hat{y}_{l+1} + z_{\epsilon/2} V_l] \quad (10.27)$$

with

$$\begin{aligned} \hat{y}_{l+1} &:= x'_{l+1} (X'_l X_l + aI_p)^{-1} X'_l Y_l, \\ V_l^2 &:= \sigma^2 x'_{l+1} (X'_l X_l + aI_p)^{-1} x_{l+1} + \sigma^2, \end{aligned} \quad (10.28)$$

and (10.26) gives the prediction interval (10.27) with

$$\begin{aligned} \hat{y}_{l+1} &:= Y'_l (K_l + aI_l)^{-1} k_l, \\ V_l^2 &:= \sigma^2 + \frac{\sigma^2}{a} \mathcal{K}(x_{l+1}, x_{l+1}) - \frac{\sigma^2}{a} k'_l (K_l + aI_l)^{-1} k_l. \end{aligned}$$

Remark We assume in this section that the variance of the noise ξ_i is known, since there are serious problems with “conjugate analysis” (the most analytically convenient variety of Bayesian learning) for the linear regression model (see, e.g., O’Hagan 1994, §9.41). This assumption will be satisfied in all our experiments.

Experimental results

In this subsection we will experimentally compare Bayesian ridge regression with the basic RRCM with the raw residuals as nonconformity scores:

$$\alpha_i = |y_i - \hat{y}_i|$$

(as on p. 30). It appears that the only realistic way to ensure that the Bayesian assumptions are satisfied is to generate the data set artificially, and this is what we do here. (Even the assumption of exchangeability is rarely satisfied for real-world data sets, but at least we have a simple and relatively nondestructive way to ensure it by permuting the data set.)

We generated a set of 506 examples with 13 attributes and one label (these numbers were chosen to make the results comparable to those obtained for the Boston Housing data set) as follows: first, a weight vector $w \in \mathbb{R}^{13}$ was generated from $N_{0,I_{13}}$, then the objects $x_1, \dots, x_{506} \in \mathbb{R}^{13}$ were independently generated from the uniform distribution on $[-1, 1]^{13}$, and finally the labels y_1, \dots, y_{506} were generated from (10.21), where the random variables ξ_i , distributed as $N_{0,1}$, were independent between themselves and of w and the objects. Therefore, the true values of a and σ are 1.

Both RRCM and Bayesian ridge regression can be said to depend on the true mechanism generating the data, but in different ways: Bayesian ridge regression uses the true mechanism directly, whereas RRCM uses it only for a rational choice of the nonconformity measure.

In Chap. 3 we considered empirical calibration and performance curves for the case of classification (Fig. 3.5 on p. 64). Similar curves for RRCM are shown in Fig. 10.1. The empirical calibration curve is defined exactly as before: for each significance level ϵ we show the percentage of errors made by the RRCM at this level on all 506 examples processed in the on-line fashion (cf. (3.16) on p. 64). The empirical performance curve is defined slightly differently: for each significance level ϵ we give the median width of prediction intervals (i.e., convex hulls of the prediction sets Γ_n^ϵ , $n = 1, \dots, 506$) at significance level ϵ , for all 506 examples. (This is similar to the approach of §2.3, see p. 39, and §8.5, but now we are interested in the dependence on ϵ rather than in the changes in the median width as new examples are processed.) As usual, the empirical calibration curve is close to the diagonal.

The picture for Bayesian ridge regression ((10.27) with (10.28)) fed with the correct parameters $a = 1$ and $\sigma = 1$ looks very similar: see Fig. 10.2.

The similarity disappears when the two algorithms are given wrong values for a . For example, let us see what happens if we tell the algorithms that the expected value of $\|w\|$ is just 1% of what it really is (this corresponds to taking $a = 10000$). The RRCM (Fig. 10.3) stays valid, but its performance deteriorates. The performance of Bayesian ridge regression (Fig. 10.4) is hardly affected, but its predictions become invalid (the empirical calibration curve deviates significantly from the diagonal, especially for the most important small significance levels). The worst that can happen to RRCM is that its

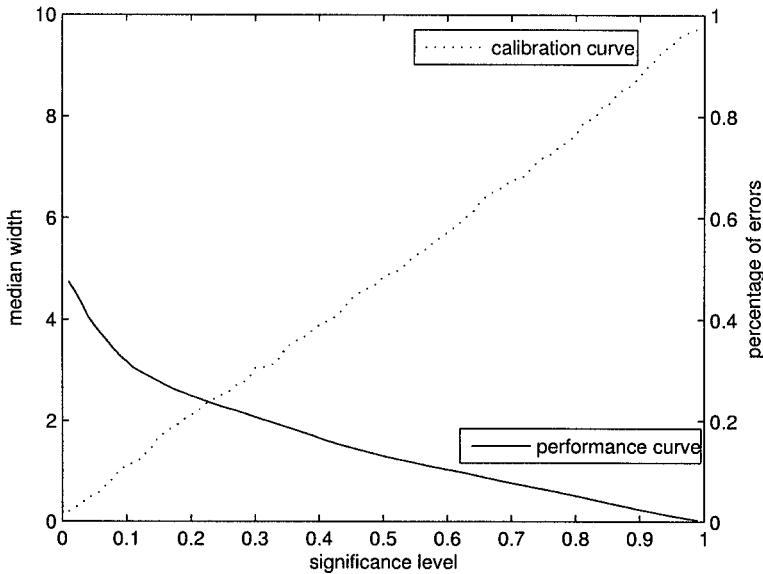


Fig. 10.1. The empirical performance (left-hand scale) and calibration (right-hand scale) curves for RRCM with $a = 1$ on the artificial data set with $a = 1$ and $\sigma = 1$

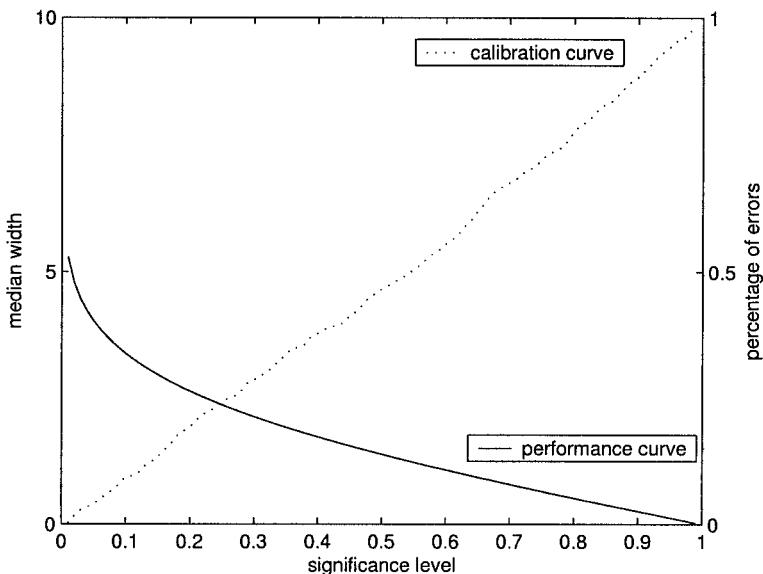


Fig. 10.2. The empirical performance (left-hand scale) and calibration (right-hand scale) curves for Bayesian ridge regression fed with $a = 1$ and $\sigma = 1$ on the artificial data set with $a = 1$ and $\sigma = 1$

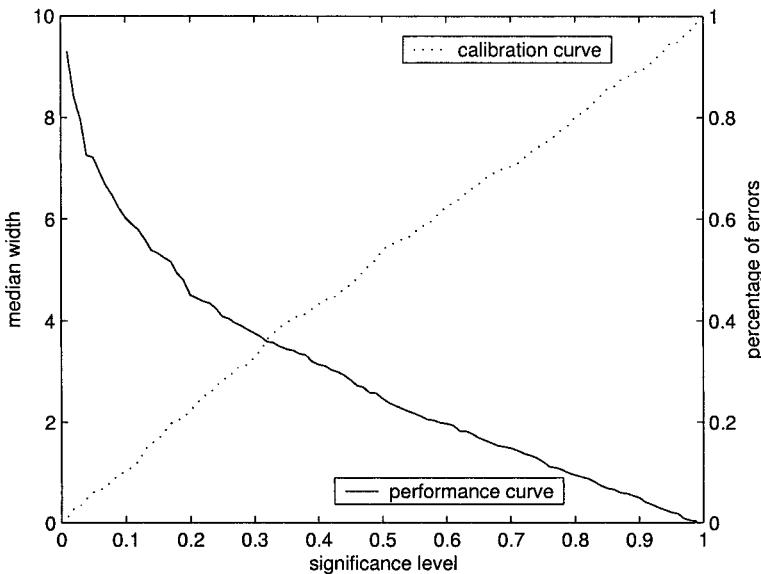


Fig. 10.3. The empirical performance (left-hand scale) and calibration (right-hand scale) curves for RRCM with $a = 10000$ on the artificial data set with $a = 1$ and $\sigma = 1$

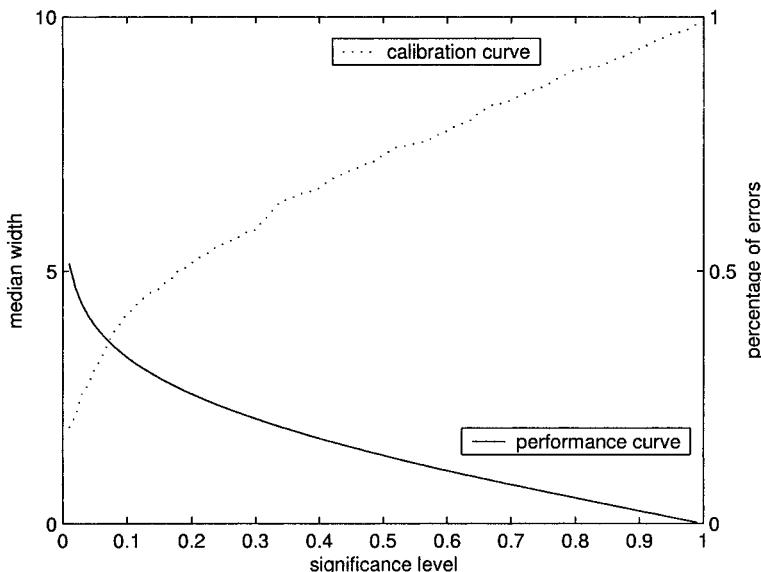


Fig. 10.4. The empirical performance (left-hand scale) and calibration (right-hand scale) curves for Bayesian ridge regression fed with $a = 10000$ and $\sigma = 1$ on the artificial data set with $a = 1$ and $\sigma = 1$

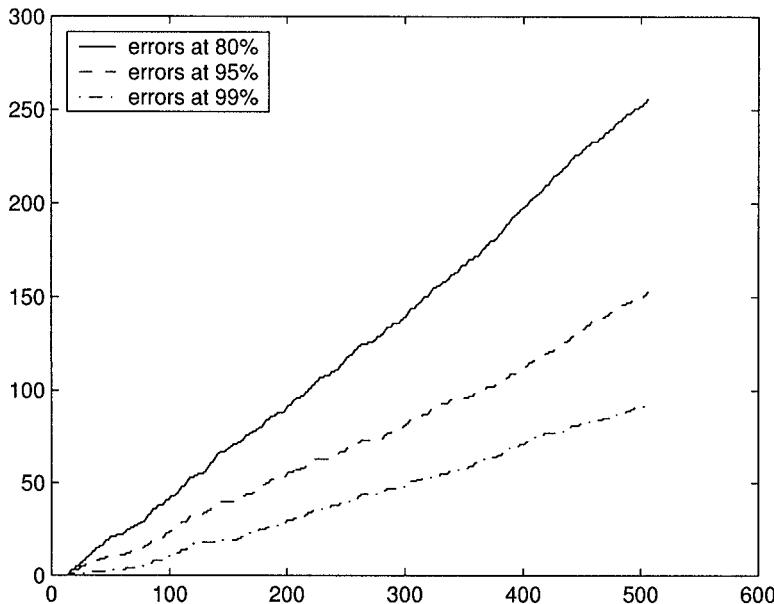


Fig. 10.5. Cumulative numbers of errors at three confidence levels for Bayesian ridge regression fed with $a = 10000$ and $\sigma = 1$ on the artificial data set with $a = 1$ and $\sigma = 1$

predictions will become useless, whereas Bayesian ridge regression's predictions can become misleading. Looking at the cumulative error lines for the latter (Fig. 10.5), we can see that they are still approximately straight but have wrong slopes.

Remark It is interesting that even Bayesian predictors based on the true probability distribution for the data do not provide the same guarantees for the validity of on-line predictions as the conformal predictors do. Indeed, suppose the true value of a is 10000. There is still a small probability that the value of the weight vector w will look as if it was generated from $N_{0,I}$ rather than $N_{0,0.0001I}$. In this case the empirical calibration curve will be like that in Fig. 10.4. This behavior, of course, will not be limited to an initial segment of the sequence of examples; there will be gross lack of validity even asymptotically.

10.4 Proofs

Proof of Proposition 10.1

Vapnik (1998, p. 341) points out that (10.13) (as well as the other results in that chapter) can be strengthened. (Equation (10.13) corresponds to Vapnik's

“Setting 2”, and we are about to describe the less restrictive “Setting 1”; for a clear discussion, see Derbeko et al. 2004, §2.1.)

Fix a bag of size $l + k$ of examples and let P be the uniform probability distribution on the set of the $(l + k)!$ orderings of that bag. Then (10.13) continues to hold if Q^{l+k} is replaced by P .

We can now apply the method used in the proof of Theorem 8.2 in §8.7. It is clear that err_n^ϵ depends on $z_1, \dots, z_{l_{n+1}}$ only through $\{z_1, \dots, z_{l_n}\}$ and $\{z_{l_n+1}, \dots, z_{l_{n+1}}\}$. Let us increase err_n^ϵ (if needed) using randomization to make the probability that $\text{err}_n^\epsilon = 1$ precisely ϵ . It is sufficient to prove that $\text{err}_1^\epsilon, \dots, \text{err}_N^\epsilon$ is a sequence of independent Bernoulli random variables. Therefore, it is sufficient to prove that $\text{err}_N^\epsilon, \dots, \text{err}_1^\epsilon$ is a sequence of independent Bernoulli random variables. This is done as before, using the fact that the conditional probability that $\text{err}_n^\epsilon = 1$ given the bag $\{z_1, \dots, z_{l_{n+1}}\}$ and the sequence $z_{l_{n+1}+1}, z_{l_{n+1}+2}, \dots$ is equal to ϵ , $n = N, \dots, 1$.

Proof of Proposition 10.2

All technical work is done in Haussler et al. 1994; we will only show how their construction fits our framework.

For any finite sequence of objects $(x_1, \dots, x_n) \in \mathbf{X}^*$, consider the following *1-inclusion graph* $G(x_1, \dots, x_n)$:

- the nodes of $G(x_1, \dots, x_n)$ are the subsets S of $\{x_1, \dots, x_n\}$ for which there exists $f \in \mathcal{F}$ such that

$$f(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise;} \end{cases}$$

- there is an edge between two nodes of $G(x_1, \dots, x_n)$ if and only if they differ in just one element $x \in \mathbf{X}$ and this element occurs in the sequence (x_1, \dots, x_n) only once.

The following is the result (stated here, as well as in the original paper, somewhat informally) which will imply Proposition 10.2.

Proposition 10.3 (Haussler et al. 1990, Lemma 2.6). *For any sequence of objects $(x_1, \dots, x_n) \in \mathbf{X}^*$, it is possible to direct all edges of $G(x_1, \dots, x_n)$ in such a way that the out-degree of each node is at most $2 \text{VC}(\mathcal{F})$ and the direction of the edges does not depend on the order in which the elements of $\{x_1, \dots, x_n\}$ appear in (x_1, \dots, x_n) .*

For each sequence (x_1, \dots, x_n) fix a directed graph $\vec{G}(x_1, \dots, x_n)$ made from $G(x_1, \dots, x_n)$ in this way.

Now we can construct the required conformal predictor. Define a nonconformity measure A_n as follows (we will omit the argument

$$\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}$$

of A_n):

- if (z_1, \dots, z_n) is not compatible with \mathcal{F} , set $A_n(z_i) := \infty$, $i = 1, \dots, n$;
- otherwise, $S := \{x_i : y_i = 1\}$ is a node of $\vec{G}(x_1, \dots, x_n)$ (where $(x_i, y_i) := z_i \in \mathbf{X} \times \mathbf{Y}$, $i = 1, \dots, n$); for each $i = 1, \dots, n$, set $A_n(z_i) := 1$ if

$$S_i := \begin{cases} S \setminus \{x_i\} & \text{if } x_i \in S \\ S \cup \{x_i\} & \text{if not} \end{cases}$$

is a node of $\vec{G}(x_1, \dots, x_n)$ and (S, S_i) is an edge of $\vec{G}(x_1, \dots, x_n)$; otherwise, set $A_n(z_i) := 0$.

To see that (10.18) (p. 261) holds, we will argue indirectly. The negation of (10.18) implies that the pair

$$\left\{ \{x_i : i \in \{1, \dots, n-1\} \& y_i = 1\}, \quad (10.29) \right.$$

$$\left. \{x_i : i \in \{1, \dots, n-1\} \& y_i = 1\} \cup \{x_n\} \right\} \quad (10.30)$$

is an edge of $G(x_1, \dots, x_n)$. If this edge is directed from (10.29) to (10.30) in $\vec{G}(x_1, \dots, x_n)$, 0 will not be included in $\Gamma_n^{\epsilon_n}(z_1, \dots, z_{n-1}, x_n)$; if it is directed from (10.30) to (10.29), 1 will not be included in $\Gamma_n^{\epsilon_n}(z_1, \dots, z_{n-1}, x_n)$. In any case, the prediction will not be multiple.

10.5 Bibliographical remarks

Inductive prediction

Our historical survey of induction follows, to a large degree, Stigler 1986a (Chap. 2). Some modern approaches to bounding the Bernoulli parameter p given a data sequence are described by Brown et al. (2001). For reviews of attempts to improve Jacob Bernoulli's $N(0.02, 1/1001, 0.6)$, see Prokhorov 1986 and Sheynin 2003. Barnard's words were quoted second-hand after Picard and Berk (1990).

Different versions of the Glivenko–Cantelli theorem were proved by Cantelli (1933), Glivenko (1933), and Kolmogorov (1933b), whose articles were published in the same volume of the same journal (for details and further developments, see Vapnik 1998, Comments and Bibliographical Remarks).

For a review of results about the VC dimension of different classes of neural networks, see Devroye et al. 1996 (§30.4) and Anthony and Bartlett 1999.

The study by Langford (2004), already mentioned in Chap. 1, found that the hold-out estimate performs much better than the sample compression bound and PAC-Bayes bounds on several data sets. These data sets are low-dimensional, but one can expect the advantage of the hold-out estimate to become even more pronounced as the number of attributes grows. It should be borne in mind that Tables 5.2 and 5.3 in that paper are not based on rigorous bounds on the true error rate of the learned classifier (as the author explains in §5.3.1).

Transductive prediction

For a good review of tolerance regions see Guttman 1970.

After the discussion of transduction had been published in Vapnik and Chervonenkis 1974 for the case of classification, it was extended by Vapnik (in the English translation of Vapnik 1982) to regression. For simplicity, Vapnik and Chervonenkis (1974) discuss only the case where the sizes of the training and working sets coincide, $l = k$; this restriction was removed by Vapnik and Sterin (1977). For a recent development of the Vapnik–Chervonenkis theory of transduction, see Derbeko et al. 2004.

Haussler et al. (1994) proved an analog of Proposition 10.2 for a smoothed version of \mathcal{F} -conformal predictor, replacing the variable significance level $2 \text{VC}(\mathcal{F})/n$ by $\text{VC}(\mathcal{F})/n$. They showed that the latter expression is optimal to within a constant factor (namely, to within a factor of 2), and Li et al. (2001) showed that it is optimal to within a factor of $1 + o(1)$.

Our philosophy of probability is described in detail in Shafer and Vovk 2001. It holds that probabilities are given an empirical interpretation by assuming that no one who gambles against them without risking bankruptcy will multiply their initial capital by a large or infinite factor.

Bayesian prediction

In §10.3 we discussed the differences between our approach and the Bayesian approach only in the case of regression. We were mainly following the paper by Melluish et al. (2001b), which also considers classification.

The theorem on normal correlation, which we used in the derivation of the kernel representation of Bayesian ridge regression, is known in geostatistics as simple kriging (see, e.g., Cressie 1993, p. 110). The name “kriging” was coined by Matheron (1963) after the South African mining engineer Krige, but the method itself is not due to Krige (Cressie 1993, p. 106).

The simulations reported in §10.3 begin with artificial data sets, since it seems very difficult to arrive at a suitable statistical model and prior for naturally occurring data sets. Experiments with the latter are attempted in Melluish et al. 2001b.

Appendix A: Probability theory

In this appendix we will give some basic definitions and results of probability theory needed for core results in this book. It is not suitable for a first study; our main goal is to familiarize the reader with our terminology and notation. The reader who needs an introduction to this material is advised to consult existing excellent textbooks such as Shiryaev 1996 and Williams 1991 (in our brief review we usually follow Kolmogorov 1933a, Shiryaev 1996, and Devroye et al. 1996). We will rarely give any proofs (and the proofs that we do give will sometimes use notions not defined and results not stated here). This appendix does not treat topics (such as linear regression and Markov chains) that are needed only for applications of this book's core results; references to the relevant literature are given in the main part of the book.

Our exposition is based on Kolmogorov's measure-theoretic axioms of probability. The recent suggestion by Shafer and Vovk (2001) to base probability theory on the theory of perfect-information games rather than measure theory would have certain advantages, but we preferred the more familiar approach.

A.1 Basics

Kolmogorov's axioms

A σ -algebra \mathcal{F} on a set Ω is a collection of subsets of Ω which contains \emptyset and Ω and is closed under the operations of complementation and taking finite and countable unions and intersections. A *measurable space* is a set Ω equipped with a σ -algebra \mathcal{F} on Ω (so formally the (Ω, \mathcal{F}) is a measurable space; we will, however, often refer to Ω as a measurable space when \mathcal{F} is clear from the context). The elements of \mathcal{F} are called *measurable sets* in Ω . A σ -algebra \mathcal{F}' such that $\mathcal{F}' \subseteq \mathcal{F}$ is called a *sub- σ -algebra* of \mathcal{F} .

For any family \mathcal{A} of subsets of a set Ω there exists the smallest σ -algebra \mathcal{F} on Ω such that $\mathcal{A} \subseteq \mathcal{F}$ (take as \mathcal{F} the intersection of all σ -algebras that

include \mathcal{A}). The smallest σ -algebra on the real line \mathbb{R} containing all intervals (a, b) is called *Borel*. Similarly, the smallest σ -algebra on $[-\infty, \infty]$ containing all intervals (a, b) and the one-element sets $\{-\infty\}$ and $\{\infty\}$ is called *Borel*. (More generally, the Borel σ -algebra on a topological space is defined as the smallest σ -algebra containing all open sets.) When considered as measurable spaces, \mathbb{R} and $[-\infty, \infty]$ will always be assumed equipped with the Borel σ -algebra.

A measurable space is *Borel* if it is isomorphic to a measurable subset of the interval $[0, 1]$. The class of Borel spaces is very rich: for example, all Polish spaces (such as finite-dimensional Euclidean spaces \mathbb{R}^n , \mathbb{R}^∞ , functional spaces C and D) are Borel; finite and countable products of Borel spaces are also Borel (see, e.g., Schervish 1995, §B.3.2).

A *probability distribution* on a measurable set (Ω, \mathcal{F}) is a function $P : \mathcal{F} \rightarrow [0, 1]$ such that: $P(\Omega) = 1$;

$$P(A \cup B) = P(A) + P(B)$$

for all disjoint $A, B \in \mathcal{F}$; and

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \rightarrow \infty} P(A_n) \quad (\text{A.1})$$

for all nested sequences $A_1 \subseteq A_2 \subseteq \dots$ of sets in \mathcal{F} .

The main object studied in probability theory is a *probability space* (Ω, \mathcal{F}, P) , where (Ω, \mathcal{F}) is a measurable space and P is a probability distribution on (Ω, \mathcal{F}) . The elements of \mathcal{F} (i.e., the measurable sets in Ω) are also called *events*. A function $\xi : \Omega \rightarrow \Xi$, where Ξ is another measurable space, is called *measurable*, or a *random element* in Ξ , if, for every measurable set $A \subseteq \Xi$, the pre-image $f^{-1}(A) \subseteq \Omega$ is measurable; we will say that ξ is \mathcal{F} -measurable if the σ -algebra on Ω has to be mentioned explicitly. Two important special cases are: random elements in \mathbb{R} are called *random variables*, and random elements in $[-\infty, \infty]$ are called *extended random variables*. The σ -algebra on Ω generated by a random element $\xi : \Omega \rightarrow \Xi$ consists of all events of the form $\xi^{-1}(A)$, A ranging over the measurable sets in Ξ . The *distribution* of a random element ξ in Ξ is the probability distribution $P\xi^{-1}$ on Ξ which is the image of P under the mapping ξ :

$$P\xi^{-1}(E) := P(\xi^{-1}(E))$$

for all events $E \subseteq \Xi$. An event E is *almost certain* if $P(E) = 1$; a property of $\omega \in \Omega$ holds *almost surely* (often abbreviated to “a.s.”) if the event that this property is satisfied is almost certain.

A *statistical model* is a family of probability distributions $(P_\theta : \theta \in \Theta)$ on the same measurable space (called the *sample space*) indexed by the elements θ of some *parameter space* Θ . Statistical models are the standard way of modeling uncertainty.

Convergence

There are many senses in which a sequence of random variables ξ_1, ξ_2, \dots can converge to a number $c \in \mathbb{R}$, but in this book we will only be interested in the following two. We say that ξ_1, ξ_2, \dots converges to c in probability if

$$\lim_{n \rightarrow \infty} P\{\omega \in \Omega : |\xi_n(\omega) - c| > \epsilon\} = 0$$

for any ϵ . The other notion of convergence is where ξ_1, ξ_2, \dots converges to c almost surely. Convergence almost surely implies convergence in probability.

A.2 Independence and products

Let (Ω, \mathcal{F}, P) be a probability space. Sub- σ -algebras $\mathcal{F}_1, \dots, \mathcal{F}_n$ of \mathcal{F} are said to be *independent* if, for any choice of events $A_1 \in \mathcal{F}_1, \dots, A_n \in \mathcal{F}_n$,

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i).$$

The sub- σ -algebras in an infinite sequence $\mathcal{F}_1, \mathcal{F}_2, \dots$ are *independent* if, for any $n = 1, 2, \dots$, the sub- σ -algebras $\mathcal{F}_1, \dots, \mathcal{F}_n$ are independent. The random elements in a sequence, finite or infinite, ξ_1, ξ_2, \dots are *independent* if the sub- σ -algebras $\mathcal{F}_1, \mathcal{F}_2, \dots$ generated by ξ_1, ξ_2, \dots , respectively, are independent.

Products of probability spaces

If $(Z_1, \mathcal{F}_1, Q_1), \dots, (Z_n, \mathcal{F}_n, Q_n)$ is a finite sequence of probability spaces, the product

$$(\Omega, \mathcal{F}, P) = \prod_{i=1}^n (Z_i, \mathcal{F}_i, Q_i)$$

is defined as follows:

- Ω is the Cartesian product $\prod_{i=1}^n Z_i$;
- \mathcal{F} is the smallest σ -algebra on Ω containing all Cartesian products $\prod_{i=1}^n A_i$, where $A_i \in \mathcal{F}_i$ for all i ;
- P is defined as the only probability distribution on (Ω, \mathcal{F}) such that

$$P\left(\prod_{i=1}^n A_i\right) = \prod_{i=1}^n Q_i(A_i),$$

for all $A_i \in \mathcal{F}_i$, $i = 1, \dots, n$.

Analogously, the product

$$(\Omega, \mathcal{F}, P) = \prod_{i=1}^{\infty} (Z_i, \mathcal{F}_i, Q_i)$$

of an infinite sequence of probability spaces $(Z_1, \mathcal{F}_1, Q_1), (Z_2, \mathcal{F}_2, Q_2), \dots$ is defined as follows:

- Ω is the Cartesian product $\prod_{i=1}^{\infty} Z_i$;
- \mathcal{F} is the smallest σ -algebra on Ω containing the Cartesian product $\prod_{i=1}^{\infty} A_i$ for every sequence $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2, \dots$ such that $A_i = \Omega_i$ from some i on;
- P is defined as the only probability distribution on (Ω, \mathcal{F}) such that

$$P \{(z_1, z_2, \dots) \in \Omega : z_i \in A_i, i = 1, \dots, n\} = \prod_{i=1}^n Q_i(A_i),$$

for all $n = 1, 2, \dots$ and all sequences $A_i \in \mathcal{F}_i, i = 1, \dots, n$.

Notice that the random variables

$$\xi_i(z_1, \dots, z_n) := z_i, \quad i = 1, \dots, n,$$

on the product $\prod_{i=1}^n (Z_i, \mathcal{F}_i, Q_i)$ are independent; the random variables

$$\xi_i(z_1, z_2, \dots) := z_i, \quad i = 1, 2, \dots,$$

on the infinite product $\prod_{i=1}^{\infty} (Z_i, \mathcal{F}_i, Q_i)$ are also independent.

Our notation for the elements of the product probability spaces will be

$$\left(\prod_{i=1}^n Z_i, \bigotimes_{i=1}^n \mathcal{F}_i, \prod_{i=1}^n Q_i \right) := \prod_{i=1}^n (Z_i, \mathcal{F}_i, Q_i)$$

(where we allow $n = \infty$), or, with the short-hand notation for measurable spaces,

$$\left(\prod_{i=1}^n Z_i, \prod_{i=1}^n Q_i, \right) := \prod_{i=1}^n (Z_i, Q_i).$$

When all the probability spaces (Z_i, Q_i) coincide, $(Z_i, Q_i) = (Z, Q)$ for all i , we will write (Z^n, Q^n) for the product $(\prod_{i=1}^n Z, \prod_{i=1}^n Q)$ and call it the n th power of (Z, Q) , with just ‘power’ meaning “ ∞ th power”.

Randomness model

Now can introduce one of the two main statistical models used in this book, the randomness model. The underlying measurable space is composed of infinite data sequences: it is the product \mathbf{Z}^{∞} , where \mathbf{Z} is the measurable space from which examples are drawn. The *randomness model* is defined to be the set of all power probability distributions Q^{∞} , Q ranging over the probability distributions on \mathbf{Z} .

A.3 Expectations and conditional expectations

Fix a probability space (Ω, \mathcal{F}, P) . If $\xi : \Omega \rightarrow \mathbb{R}$ is a nonnegative random variable and $A \in \mathcal{F}$ is an event, the Lebesgue integral $\int_A \xi(\omega) P(d\omega)$ is defined to be

$$\lim_{\lambda \downarrow 0} \sum_{k=0}^{\infty} k\lambda P\{\omega \in A : k\lambda \leq \xi(\omega) < (k+1)\lambda\}$$

(this limit always exists but can be infinite). If $\xi \geq 0$ is an extended random variable, the integral is defined in the same way if $P(A \cap \{\xi = \infty\}) = 0$ and is defined to be ∞ otherwise. For an arbitrary (extended) random variable ξ we set

$$\int_A \xi(\omega) P(d\omega) := \int_A \xi^+(\omega) P(d\omega) - \int_A \xi^-(\omega) P(d\omega),$$

provided at least one of $\int_A \xi^+(\omega) P(d\omega)$, $\int_A \xi^-(\omega) P(d\omega)$ is finite; if both are infinite, $\int_A \xi(\omega) P(d\omega)$ does not exist. A shorter alternative notation for $\int_A \xi(\omega) P(d\omega)$ is $\int_A \xi dP$, with $\int_{\Omega} \xi dP$ abbreviated to $\int \xi dP$. When the probability space (Ω, \mathcal{F}, P) is clear from the context, we will write $\mathbb{E} \xi$ for $\int \xi dP$; $\mathbb{E}_P \xi$ is synonymous with $\int \xi dP$. (Similarly, we will sometimes write $\mathbb{P}(A)$ or $\mathbb{P}_P(A)$ for $P(A)$.)

Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub- σ -algebra of \mathcal{F} and ξ be a nonnegative extended random variable. The conditional expectation $\mathbb{E}(\xi | \mathcal{G})$ of ξ w.r. to \mathcal{G} is defined to be a \mathcal{G} -measurable extended random variable such that, for any $A \in \mathcal{G}$,

$$\int_A \xi(\omega) P(d\omega) = \int_A \mathbb{E}(\xi | \mathcal{G})(\omega) P(d\omega)$$

(any two \mathcal{G} -measurable extended random variables satisfying this condition coincide almost surely; they will be referred to as *versions* $\mathbb{E}(\xi | \mathcal{G})$). For general extended random variables ξ , define

$$\mathbb{E}(\xi | \mathcal{G}) := \mathbb{E}(\xi^+ | \mathcal{G}) - \mathbb{E}(\xi^- | \mathcal{G});$$

this definition will be used only when

$$\min(\mathbb{E}(\xi^+ | \mathcal{G}), \mathbb{E}(\xi^- | \mathcal{G})) < \infty \quad \text{a.s.}$$

There may be many *versions* of $\mathbb{E}(\xi | \mathcal{G})$, but any two of them coincide almost surely.

In this book we use the following properties of conditional expectations (for proofs of the first two of these see, e.g., Shiryaev 1996, §II.7.4; the third is obvious):

1. If \mathcal{G} is a sub- σ -algebra of \mathcal{F} , ξ and η are bounded random variables, and η is \mathcal{G} -measurable,

$$\mathbb{E}(\eta \xi | \mathcal{G}) = \eta \mathbb{E}(\xi | \mathcal{G})$$

almost surely.

2. If \mathcal{G}_1 and \mathcal{G}_2 are sub- σ -algebras of \mathcal{F} , $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$, and ξ is a random variable,

$$\mathbb{E}(\mathbb{E}(\xi | \mathcal{G}_2) | \mathcal{G}_1) = \mathbb{E}(\xi | \mathcal{G}_1)$$

almost surely; in particular, for any sub- σ -algebra \mathcal{G} of \mathcal{F} ,

$$\mathbb{E}(\mathbb{E}(\xi | \mathcal{G})) = \mathbb{E}(\xi).$$

3. If \mathcal{G} is a sub- σ -algebra of \mathcal{F} , ξ is a nonnegative random variable, $(P_\theta : \theta \in \Theta)$ is a statistical model with Θ a measurable space such that the function $\theta \mapsto P_\theta(E)$ is measurable for any event E , μ is a probability distribution on Θ , and $P = \int P_\theta \mu(d\theta)$ is the mixture of the probability distributions P_θ , then

$$\mathbb{E}_P(\xi | \mathcal{G}) = \int \mathbb{E}_{P_\theta}(\xi | \mathcal{G}) \mu(d\theta)$$

almost surely (the underlying probability distribution is given as a lower index).

For an event $A \in \mathcal{F}$ and a sub- σ -algebra $\mathcal{G} \subseteq \mathcal{F}$, the *conditional probability* $\mathbb{P}(A | \mathcal{G})$ is defined as the conditional expectation $\mathbb{P}(\mathbb{I}_A | \mathcal{G})$ of A 's indicator function. We sometimes write $\mathbb{E}(\xi | \eta)$ (resp. $\mathbb{P}(A | \eta)$), where η is a random element, to mean $\mathbb{E}(\xi | \mathcal{G})$ (resp. $\mathbb{P}(A | \mathcal{G})$) where \mathcal{G} is the σ -algebra generated by η .

A.4 Markov kernels and regular conditional distributions

Let Ω and Z be two measurable spaces. A function $Q(\omega, A)$, usually written as $Q(A | \omega)$, where ω ranges over Ω and A over the measurable sets in Z , is called a *Markov kernel* if:

- as a function of A , $Q(A | \omega)$ is a probability distribution on Z , for each $\omega \in \Omega$;
- as a function of ω , $Q(A | \omega)$ is measurable, for each measurable $A \subseteq Z$.

We will say that Q is a Markov kernel of the type $\Omega \hookrightarrow Z$, using \hookrightarrow to distinguish Markov kernels from functions of the type $\Omega \rightarrow Z$. Unlike functions, Markov kernels map $\omega \in \Omega$ to probability distributions on Z ; we will sometimes use the notation $Q(\omega)$ for the probability distribution $A \mapsto Q(A | \omega)$ on Z and write $Q(\omega)(d\zeta)$ for $Q(d\zeta | \omega)$. If $E \subseteq \Omega$ is an event in Ω , the restriction $Q|_E$ is defined to be the same Markov kernel $Q(A | \omega)$ but with ω ranging over E .

Lemma A.1. *If $Q : \Omega \hookrightarrow Z$ is a Markov kernel and a function f on $\Omega \times Z$ is measurable, the function $\omega \in \Omega \mapsto \int f(\omega, \zeta) Q(d\zeta | \omega)$ is also measurable.*

Proof. The statement of the lemma follows from the standard monotone-class argument (see, e.g., Williams 1991, §3.14) and the fact that it holds for the indicator functions $f = \mathbb{I}_{A \times B}$ of the rectangles $A \times B$, where $A \subseteq \Omega$ and $B \subseteq Z$ are measurable. \square

Regular conditional distributions

Let (Ω, \mathcal{F}, P) be a probability space. If we fix $\omega \in \Omega$, $\mathbb{P}(A | \mathcal{G})(\omega)$ will not necessarily be a probability distribution as a function of $A \in \mathcal{F}$. We cannot even guarantee that $\mathbb{P}(A | \mathcal{G})(\omega)$ will be a probability distribution for almost all ω . Consider, e.g., the property

$$\mathbb{P}(A \cup B | \mathcal{G})(\omega) = \mathbb{P}(A | \mathcal{G})(\omega) + \mathbb{P}(B | \mathcal{G})(\omega),$$

where A and B are disjoint events. For fixed A and B it will hold for almost all ω , but the set of exceptional ω for which it does not hold will depend on the pair (A, B) , and the union of the exceptional sets over the potentially uncountable set of all pairs (A, B) may not be an exceptional set.

Let \mathcal{G}_1 and \mathcal{G}_2 be two sub- σ -algebras of \mathcal{F} . A Markov kernel

$$Q : (\Omega, \mathcal{G}_1) \hookrightarrow (\Omega, \mathcal{G}_2) \tag{A.2}$$

is called a *regular conditional probability* if for each $A \in \mathcal{G}_2$, $Q(A | \omega)$ as a function of ω is a version of the conditional probability $\mathbb{P}(A | \mathcal{G}_1)(\omega)$. There may be additional regularity properties that one might like to impose, such as

$$\mathbb{P}\{\omega : \forall A \in \mathcal{G}_1 \cap \mathcal{G}_2 : Q(A | \omega) = \mathbb{I}_A(\omega)\} = 1,$$

but they do not form part of the definition.

The two most standard general results about existence of regular conditional probabilities are:

- A regular conditional probability exists if Ω is a Lusin space (i.e., a topological space that is homeomorphic to a Borel subset of a compact metric space) and \mathcal{G}_1 is the Borel σ -algebra. (This condition does not depend on \mathcal{G}_2 , so one may take $\mathcal{G}_2 := \mathcal{F}$.)
- A regular conditional probability exists if \mathcal{G}_2 is the σ -algebra generated by a random variable. (It is clear that “random variable” can be replaced by “random element with values in a Borel space”.)

Proofs can be found in, e.g., Rogers and Williams 1994 (§II.89) for the first statement and Shiryaev 1996 (Theorem II.7.5) for the second.

The following lemma asserts that, as one would expect, conditional expectations can be computed by averaging over regular conditional probabilities. We will need it only for the case $\mathcal{G}_2 = \mathcal{F}$, but even the more general statement is very easy to prove (and so we will give a proof for this standard result, which, however, can rarely be found in modern textbooks).

Lemma A.2. *Let Q be a Markov kernel from (Ω, \mathcal{G}_1) to (Ω, \mathcal{G}_2) , where \mathcal{G}_1 and \mathcal{G}_2 are sub- σ -algebras of \mathcal{F} . The Markov kernel Q is a regular conditional probability if and only if, for any bounded \mathcal{G}_2 -measurable random variable ξ , $\omega_1 \in \Omega \mapsto \int \xi(\omega_2)Q(d\omega_2 | \omega_1)$ is a version of $\mathbb{E}(\xi | \mathcal{G}_1)$.*

Proof. We are required to prove that Q is a regular conditional probability if and only if

$$\int_E \int \xi(\omega_2) Q(d\omega_2 | \omega_1) P(d\omega_1) = \int_E \xi(\omega) P(d\omega) \quad (\text{A.3})$$

for all \mathcal{G}_1 -measurable E and all bounded \mathcal{G}_2 -measurable ξ .

If we take $\xi = \mathbb{I}_A$ for $A \in \mathcal{G}_2$, (A.3) will become

$$\int_E Q(A | \omega_1) P(d\omega_1) = P(E \cap A), \quad (\text{A.4})$$

which means that $Q(A | \omega)$ is a version of the conditional probability of A given \mathcal{G}_1 and, therefore, that Q is a regular conditional probability.

Let us now assume that Q is a regular conditional probability; we are required to prove that (A.3) holds for all \mathcal{G}_1 -measurable E and bounded \mathcal{G}_2 -measurable ξ ; fix such an E . We know that (A.3) holds for $\xi = \mathbb{I}_A$ with $A \in \mathcal{G}_2$ (see (A.4)); it remains to apply the standard monotone-class argument (Williams 1991, §3.14; already used in the proof of Lemma A.1). \square

If $\xi_1 : \Omega \rightarrow \Xi_1$ and $\xi_2 : \Omega \rightarrow \Xi_2$ are random elements, a *regular conditional distribution* of ξ_2 given ξ_1 is the Markov kernel $R : (\Omega, \mathcal{G}_1) \hookrightarrow \Xi_2$ defined by $R(\omega) := Q(\omega) \xi^{-1}$, $\omega \in \Omega$, where \mathcal{G}_1 is the σ -algebra on Ω generated by ξ_1 and Q is a regular conditional probability (A.2); a regular conditional distribution R exists if and only if a regular conditional probability (A.2) exists.

A.5 Exchangeability

Let \mathbf{Z} be a measurable space. We say that a probability distribution P on the measurable space \mathbf{Z}^n of sequences of length n , where $n \in \{1, 2, \dots\}$, is *exchangeable* if

$$P(E) = P\{z_1, \dots, z_n : z_{\pi(1)} \dots z_{\pi(n)} \in E\}$$

for any measurable $E \subseteq \mathbf{Z}^n$ and any permutation π of the set $\{1, \dots, n\}$ (in words: if the distribution of the sequence $z_1 \dots z_n$ is invariant under any permutation of the indices). We say that a probability distribution P on the power measurable space \mathbf{Z}^∞ is *exchangeable* if the marginal distribution P_n of P on \mathbf{Z}^n (defined by

$$P_n(E) := P\{(z_1, z_2, \dots) \in \mathbf{Z}^\infty : z_1 \dots z_n \in E\}$$

for all events $E \subseteq \mathbf{Z}^n$) is exchangeable for each $n = 1, 2, \dots$ (in words: if the distribution of the sequence $z_1 z_2 \dots$ is invariant under any permutation of a finite number of the indices). The *exchangeability model* is defined to be the set of all exchangeable probability distributions on \mathbf{Z}^∞ .

The question of relation between randomness and exchangeability models is a popular topic in the foundations of Bayesian statistics (see, e.g., Chap. 1 of Schervish 1995). Every power probability distribution Q^∞ on \mathbf{Z}^∞ is exchangeable, and under a weak regularity condition every exchangeable probability distribution on \mathbf{Z}^∞ is a mixture of power distributions; this is de Finetti's representation theorem (see, e.g., Schervish 1995, Theorem 1.49).

De Finetti's theorem. *Suppose \mathbf{Z} is a Borel space. A probability distribution P on \mathbf{Z}^∞ is exchangeable if and only if P is a mixture of power distributions:*

$$P = \int Q^\infty \mu(dQ)$$

for some probability distribution μ on the space $\mathbf{P}(\mathbf{Z})$ of all probability distributions on \mathbf{Z} (equipped with the smallest σ -algebra such that all evaluation functions $Q \mapsto Q(E)$ are measurable, E ranging over the events in \mathbf{Z}).

Conditional probabilities given a bag

We will need the following simple result about the existence of a regular conditional probability for exchangeable distributions. We define the *bag σ -algebra* on \mathbf{Z}^n as the family of events $E \subseteq \mathbf{Z}^n$ such that

$$(z_1, \dots, z_n) \in E \implies (z_{\pi(1)}, \dots, z_{\pi(n)}) \in E$$

for all permutations π of the set $\{1, \dots, n\}$.

Lemma A.3. *Let P be an exchangeable distribution on \mathbf{Z}^n for an $n \in \{1, 2, \dots\}$ and let \mathcal{G} be the bag σ -algebra. The Markov kernel C which maps each $\omega = (z_1, \dots, z_n) \in \mathbf{Z}^n$ to the probability distribution $C(\omega)$ on \mathbf{Z}^n concentrated on the set of all permutations $(z_{\pi(1)}, \dots, z_{\pi(n)})$ and assigning the same probability $1/n!$ to each of these permutations will be a regular conditional probability w.r. to \mathcal{G} in the probability space (\mathbf{Z}^n, P) .*

Proof. For any random variable ξ on \mathbf{Z}^n set

$$\bar{\xi}(z_1, \dots, z_n) := \frac{1}{n!} \sum_{\pi} \xi(z_{\pi(1)}, \dots, z_{\pi(n)}),$$

the sum being over all $n!$ permutations π of $\{1, \dots, n\}$. By Lemma A.2, we are required to prove that

$$\int_E (\bar{\xi}(z_1, \dots, z_n) - \xi(z_1, \dots, z_n)) P(dz_1, \dots, dz_n) = 0 \quad (\text{A.5})$$

for any set $E \subseteq \mathbf{Z}^n$ in the bag σ -algebra.

First we notice that, if Ω is a measurable space and $G : \Omega \rightarrow \Omega$ is a bijection measurable in both directions, then for every measurable function $f : \Omega \rightarrow \mathbb{R}$, measurable set $E \subseteq \Omega$, and probability distribution P on Ω ,

$$\int_E f dP = \int_{E'} f' dP', \quad (\text{A.6})$$

where the set E' , function f' , and probability distribution P' are defined by

$$E' := G^{-1}(E), \quad f'(\omega) := f(G(\omega)), \quad P'(A) := P(G(A)).$$

Applying this to $\Omega := \mathbf{Z}^n$, $G(z_1, \dots, z_n) := (z_{\pi(1)}, \dots, z_{\pi(n)})$, where π is a permutation, and $f := \xi$, we obtain

$$\int_E (\xi(z_{\pi(1)}, \dots, z_{\pi(n)}) - \xi(z_1, \dots, z_n)) P(dz_1, \dots, dz_n) = 0$$

(remember that $E' = E$ and $P' = P$). Finally, averaging over all π gives (A.5). \square

A.6 Theory of martingales

Let (Ω, \mathcal{F}, P) be a probability space. It will be convenient to use the adjectives “increasing” and “decreasing” in the extended sense (as we do throughout the book: see p. 3); in particular, a sequence of σ -algebras \mathcal{F}_n is *increasing* if $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ and a sequence of random variables ξ_n is *increasing* if $\xi_1 \leq \xi_2 \leq \dots$ ($\xi \leq \eta$ can be defined as “ $\xi(\omega) \leq \eta(\omega)$ for all ω ” or as “ $\xi \leq \eta$ almost surely”; it does not matter which definition is used for the mathematical results stated in this appendix).

A *filtration* is an increasing sequence of sub- σ -algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}$; when we say that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}$ is a filtration, we always mean that it is complemented by $\mathcal{F}_0 := \{\emptyset, \Omega\}$. Let \mathcal{F}_∞ be the smallest σ -algebra containing all \mathcal{F}_n . We say that a sequence of random elements ξ_n , where $n = 1, 2, \dots$ or $n = 0, 1, 2, \dots$, is *adapted* if each ξ_n is \mathcal{F}_n -measurable. A sequence of random elements ξ_1, ξ_2, \dots is *predictable* if each ξ_n is \mathcal{F}_{n-1} -measurable.

We say that an adapted sequence of random variables ξ_0, ξ_1, \dots is a *martingale* if $\mathbb{E}(\xi_n | \mathcal{F}_{n-1}) = \xi_{n-1}$ for all $n = 1, 2, \dots$, and we say that an adapted sequence of random variables ξ_1, ξ_2, \dots is a *martingale difference* if $\mathbb{E}(\xi_n | \mathcal{F}_{n-1}) = 0$ for all $n = 1, 2, \dots$. A very useful generalization of the notion of a martingale is that of a *supermartingale*: this is an adapted sequence of random variables ξ_0, ξ_1, \dots such that $\mathbb{E}(\xi_n | \mathcal{F}_{n-1}) \leq \xi_{n-1}$ for all $n = 1, 2, \dots$.

If there is no *a priori* fixed filtration on the given probability space, we say that an adapted sequence of random variables ξ_0, ξ_1, \dots is a *martingale* (resp. *supermartingale*) if it is a martingale (resp. supermartingale) w.r. to the filtration \mathcal{F}_n , $n = 0, 1, \dots$, such that each \mathcal{F}_n is generated by the random variables ξ_1, \dots, ξ_n (in particular, $\mathcal{F}_0 := \{\emptyset, \Omega\}$). This rather old-fashioned notion of a martingale is used in Chap. 7.

Analogous definitions can be given for a finite filtration, $\mathcal{F}_1, \dots, \mathcal{F}_N$; in this case, martingales and supermartingales are finite sequences $\xi_0, \xi_1, \dots, \xi_N$, and predictable sequences and martingale differences are finite sequences ξ_1, \dots, ξ_N .

The following is a simple version of Doob's inequality; it holds for both finite and infinite filtrations.

Doob's inequality. *If (ξ_n) is a nonnegative supermartingale w.r. to a filtration (\mathcal{F}_n) with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and C is a positive constant, then*

$$\mathbb{P} \left\{ \sup_n \xi_n \geq C \right\} \leq \frac{\xi_0}{C} .$$

The next result says that the probability of an event is the infimum of the initial values of nonnegative martingales (or of nonnegative supermartingales) that exceed or equal one whenever the event happens; for a proof, see Shafer and Vovk 2001 (§8.5).

Ville's theorem. *If $\mathcal{F}_1, \mathcal{F}_2, \dots$ is a filtration, $\mathcal{F}_0 := \{\emptyset, \Omega\}$, and $E \in \mathcal{F}_\infty$, then*

$$\mathbb{P}(E) = \inf \left\{ \xi_0 : \liminf_{n \rightarrow \infty} \xi_n \geq \mathbb{I}_E \right\} = \inf \left\{ \xi_0 : \sup_n \xi_n \geq \mathbb{I}_E \right\}, \quad (\text{A.7})$$

where (ξ_n) ranges over the nonnegative martingales (or supermartingales) w.r. to the filtration (\mathcal{F}_n) .

In stating the following simple but useful result we will use the logical notation $A \Leftrightarrow B$ for the symmetric difference of events A and B .

Borel–Cantelli–Lévy lemma. *If $\mathcal{F}_0, \mathcal{F}_1, \dots$ is a filtration and $A_n \in \mathcal{F}_n$ for $n = 1, 2, \dots$, then*

$$\left(\sum_{n=1}^{\infty} \mathbb{P}(A_n | \mathcal{F}_{n-1}) < \infty \right) \Leftrightarrow \left(\sum_{n=1}^{\infty} \mathbb{I}_{A_n} < \infty \right)$$

almost surely.

The Borel–Cantelli–Lévy lemma generalizes the part of the classical Borel–Cantelli lemma that deals with sequences of independent events A_n (the independence of the events A_n means that the σ -algebras in the sequence

$$\mathcal{F}_n := \{\emptyset, A_n, \Omega \setminus A_n, \Omega\}$$

are independent).

Borel–Cantelli lemma. *Let A_1, A_2, \dots be a sequence of events.*

- If $\sum_n \mathbb{P}(A_n) < \infty$,

$$\sum_n \mathbb{I}_{A_n} < \infty \quad a.s.$$

- If the events A_1, A_2, \dots are independent, and $\sum_n \mathbb{P}(A_n) = \infty$,

$$\sum_n \mathbb{I}_{A_n} = \infty \quad a.s.$$

Limit theorems

In this subsection we will state some fundamental limit theorems of the theory of martingales (for proofs, see Shiryaev 1996 and Shafer and Vovk 2001).

Martingale strong law of large numbers. Let (ξ_n) be a martingale difference w.r. to a filtration $\mathcal{F}_0, \mathcal{F}_1, \dots$ and let (A_n) be an increasing predictable sequence w.r. to the same filtration with $A_1 > 0$ and $A_\infty = \infty$ a.s. If

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}(\xi_i^2 | \mathcal{F}_{i-1})}{A_i^2} < \infty \quad a.s. ,$$

then

$$\frac{1}{A_n} \sum_{i=1}^n \xi_i \rightarrow 0 \quad (n \rightarrow \infty) \quad a.s.$$

These are important special cases.

Kolmogorov's strong law of large numbers. Suppose ξ_1, ξ_2, \dots is a sequence of independent zero-mean random variables and A_1, A_2, \dots is an increasing sequence of positive numbers such that $A_n \rightarrow \infty$ ($n \rightarrow \infty$). If

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}(\xi_i^2)}{A_i^2} < \infty , \tag{A.8}$$

then

$$\frac{1}{A_n} \sum_{i=1}^n \xi_i \rightarrow 0 \quad (n \rightarrow \infty) \quad a.s.$$

Borel's strong law of large numbers. If ξ_1, ξ_2, \dots is a sequence of independent binary (i.e., taking values in $\{0, 1\}$) random variables with $\mathbb{E}(\xi_n) = p$, $n = 1, 2, \dots$, then

$$\frac{1}{n} \sum_{i=1}^n \xi_i \rightarrow p \quad (n \rightarrow \infty) \quad a.s.$$

The following result is a martingale version (due to Stout 1970) of Kolmogorov's law of the iterated logarithm (it uses the usual logical convention that the event $A \Rightarrow B$ is the union of B and the complement of A).

Martingale law of the iterated logarithm. Let (ξ_n) be a martingale difference w.r. to a filtration $\mathcal{F}_0, \mathcal{F}_1, \dots$ and let $(A_n), (c_n)$ be increasing positive predictable sequences w.r. to this filtration. If $|\xi_n| \leq c_n$ for all n , then, almost surely,

$$\left(A_n \rightarrow \infty \text{ & } c_n = o\left(\sqrt{\frac{A_n}{\ln \ln A_n}}\right) \right) \implies \limsup_{n \rightarrow \infty} \frac{\sum_{i=1}^n \xi_i}{\sqrt{2A_n \ln \ln A_n}} = 1 .$$

A.7 Hoeffding's inequality and McDiarmid's theorem

Hoeffding's inequality. Let $\mathcal{F}_0, \dots, \mathcal{F}_n$ be a filtration. For any deterministic sequence c_1, \dots, c_n of positive numbers, any predictable sequence v_1, \dots, v_n w.r. to (\mathcal{F}_i) , any martingale difference ξ_1, \dots, ξ_n w.r. to (\mathcal{F}_i) such that $|\xi_i - v_i| \leq c_i$, $i = 1, \dots, n$, and any $\epsilon > 0$,

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n \xi_i \geq \epsilon \right\} \leq \exp\left(-\frac{\epsilon^2 n^2}{2 \sum_{i=1}^n c_i^2}\right) \quad (\text{A.9})$$

and

$$\mathbb{P}\left\{ \frac{1}{n} \sum_{i=1}^n \xi_i \leq -\epsilon \right\} \leq \exp\left(-\frac{\epsilon^2 n^2}{2 \sum_{i=1}^n c_i^2}\right) . \quad (\text{A.10})$$

The proof of this result will easily follow from the following elementary but basic fact.

Lemma A.4. Let ξ be a random variable such that $\mathbb{E} \xi = 0$ and $a \leq \xi \leq b$ for constants a and b and let $s > 0$. Then

$$\mathbb{E}(e^{s\xi}) \leq e^{s^2(b-a)^2/8} .$$

Proof. Assume, without loss of generality, that $s = 1$ (the general case follows from this special case by replacing ξ, a, b with $s\xi, sa, sb$, respectively). Since the function $x \mapsto e^x$ is convex, we can assume, without loss of generality, that the distribution of ξ is concentrated on $\{a, b\}$. Since $\mathbb{E} \xi = 0$, the mass at a is $b/(b-a)$ and the mass at b is $-a/(b-a)$ (remember that $a \leq 0$ and $b \geq 0$; we only consider the nontrivial case $a \neq b$); therefore, we are only required to prove that

$$\frac{b}{b-a} e^a - \frac{a}{b-a} e^b \leq e^{(b-a)^2/8} . \quad (\text{A.11})$$

(The formal version of this argument is that, by the convexity of the exponential function,

$$e^x \leq \frac{x-a}{b-a} e^b + \frac{b-x}{b-a} e^a$$

for $x \in [a, b]$; it remains to find the expectations of the two sides.)

Introducing the notation

$$u := b - a, \quad p := -\frac{a}{b-a}, \quad 1-p := \frac{b}{b-a},$$

we can rewrite (A.11) as

$$pe^{(1-p)u} + (1-p)e^{-pu} \leq e^{u^2/8}$$

or, equivalently,

$$\phi(u) := -pu + \ln(1 - p + pe^u) \leq u^2/8.$$

It remains to notice that $\phi(0) = 0$, $\phi'(0) = 0$ and $\phi''(u) \leq 1/4$; the last inequality follows from the fact that the geometric mean never exceeds the arithmetic mean:

$$\phi''(u) = \left(-p + \frac{p}{p + (1-p)e^{-u}} \right)' = \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \leq \frac{1}{4}. \quad \square$$

Lemma A.4 shows that

$$\exp \left(s \sum_{i=1}^n \xi_i - \frac{1}{8} s^2 \sum_{i=1}^n (b_i - a_i)^2 \right), \quad n = 0, 1, 2, \dots, \quad (\text{A.12})$$

where a_i and b_i are predictable sequences and $\xi_i \in [a_i, b_i]$ is a martingale difference, is a supermartingale; this fact is used directly in Chaps. 6 and 7. In conjunction with Doob's inequality it implies Hoeffding's inequality (s should be chosen optimally for the given ϵ and c_1, \dots, c_n).

McDiarmid's theorem. *Let $n \in \mathbb{N}$, ξ_1, \dots, ξ_n be a sequence of independent random elements taking values in a measurable space A , and a measurable function $f : A^n \rightarrow \mathbb{R}$ satisfy*

$$\sup_{x_1, \dots, x_n, x'_i \in A} |f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_n)| \leq c_i, \\ i = 1, \dots, n,$$

for some deterministic sequence c_1, c_2, \dots . For any $\epsilon > 0$,

$$\mathbb{P}\left\{ f(\xi_1, \dots, \xi_n) - \mathbb{E} f(\xi_1, \dots, \xi_n) \geq \epsilon \right\} \leq \exp\left(-2\epsilon^2 / \sum_{i=1}^n c_i^2\right)$$

and

$$\mathbb{P}\left\{ \mathbb{E} f(\xi_1, \dots, \xi_n) - f(\xi_1, \dots, \xi_n) \geq \epsilon \right\} \leq \exp\left(-2\epsilon^2 / \sum_{i=1}^n c_i^2\right).$$

Proof. Let \mathcal{F}_i be the σ -algebra generated by ξ_1, \dots, ξ_i , $i = 0, 1, \dots, n$ (and so $\mathcal{F}_0 = \{\emptyset, \Omega\}$). Assume, without loss of generality, that $\mathbb{E} f(\xi_1, \dots, \xi_n) = 0$. In this proof we will use the notation

$$\begin{aligned}\text{ess inf}(\zeta | \mathcal{F}) &:= \sup \{c \in \mathbb{R} : \mathbb{P}(\{\zeta \geq c\} | \mathcal{F}) = 1\}, \\ \text{ess sup}(\zeta | \mathcal{F}) &:= \inf \{c \in \mathbb{R} : \mathbb{P}(\{\zeta \leq c\} | \mathcal{F}) = 1\}.\end{aligned}$$

Set

$$\begin{aligned}\eta_i &:= \mathbb{E}(f(\xi_1, \dots, \xi_n) | \mathcal{F}_i), \quad i = 0, 1, \dots, n, \\ \Delta\eta_i &:= \eta_i - \eta_{i-1}, \quad i = 1, \dots, n, \\ v_i &:= \frac{1}{2} (\text{ess inf}(\Delta\eta_i | \mathcal{F}_{i-1}) + \text{ess sup}(\Delta\eta_i | \mathcal{F}_{i-1})), \quad i = 1, \dots, n.\end{aligned}$$

Noticing that

$$|\Delta\eta_i - v_i| \leq c_i/2, \quad i = 1, \dots, n,$$

it remains to apply Hoeffding's inequality to the deterministic sequence $c_i/2$, the predictable sequence v_i , and the martingale difference $\Delta\eta_i$. \square

A.8 Bibliographical remarks

Kolmogorov's axioms are proposed in Kolmogorov 1933a.

Conditional probabilities

The classical definition of the conditional probability of an event A given another event B is $\mathbb{P}(A|B) := \mathbb{P}(A \cap B)/\mathbb{P}(B)$, but it only works if $\mathbb{P}(B) > 0$. One of the main contributions of Kolmogorov's *Grundbegriffe* (1933a) was to extend this definition to the case where $\mathbb{P}(B) = 0$ is allowed; the price was that the definition had to be given for all B in a partition of Ω simultaneously, and the definition made sense not always but only almost surely. From the modern point of view, presented in §A.3, Kolmogorov defined $\mathbb{P}(A|\mathcal{G})$ only for \mathcal{G} obtained from the original σ -algebra \mathcal{F} and a partition: an event $A \in \mathcal{F}$ is included in \mathcal{G} if and only if it is the union of some elements of the partition; his definition, however, extends trivially to the standard definition (given above) applicable to an arbitrary σ -algebra $\mathcal{G} \subseteq \mathcal{F}$. (For a detailed discussion, see Shafer and Vovk 2003.)

One of the difficulties of Kolmogorov's definition is demonstrated by Dieudonné's (1948) famous example, in which the conditional probabilities $\mathbb{P}(A|\mathcal{G})(\omega)$ do not have versions that would form a probability distribution as a function of A for almost all ω . This prompted the development of the theory of regular conditional probabilities, in which conditional probabilities $\mathbb{P}(A|B)$ are defined simultaneously for all events $A \in \mathcal{G}_2$ and $B \in \mathcal{G}_1$ ranging over sub- σ -algebras \mathcal{G}_2 and \mathcal{G}_1 of \mathcal{F} .

Martingales

A mathematical notion of a martingale was introduced explicitly and used for the purposes of the foundations of probability by Ville (1939); earlier, it had been used by several people, including Lévy and Kolmogorov, without an explicit definition. The simple version of Doob's inequality given in this appendix was proved already by Ville (1939); a more sophisticated version appeared in Doob 1953. A special case of Ville's theorem was proved by Ville (1939). For further historical details and references, see Shafer and Vovk 2001.

Hoeffding's inequality and McDiarmid's theorem

Hoeffding's inequality was first published in Hoeffding 1963; its martingale version is sometimes referred to as the Hoeffding–Azuma inequality (after Azuma's paper 1967), although already Hoeffding 1963 (p. 18) contains the martingale extension. Our exposition follows Devroye et al. 1996. In many textbooks Hoeffding's inequality is given in the simplified and easier to prove form with $v_i = 0$, but this simplified form does not allow obtaining the best constants in McDiarmid's theorem (McDiarmid 1989).

Appendix B: Data sets

In this appendix we will describe the two main data sets used in the book. A separate section, §B.3, is devoted to the important problem of “normalization” of the objects; some other technical issues are discussed in §B.4.

B.1 USPS data set

The USPS (US Postal Service) data set is a standard benchmark for testing classification algorithms. It consists of 7291 training examples and 2007 test examples collected from real-life US zip codes (mail passing through the Buffalo, NY, post office). In our experiments, we always merge the training set and the test set, in this order, obtaining what we call the *full USPS data set* (or just USPS data set). After that we often apply a random permutation.

Each example consists of an image (16×16 matrix with entries in the interval $(-1, 1)$ that describe the brightness of individual pixels) and its label (0 to 9).

It is well known that the USPS data set is heterogeneous; in particular, the training and test sets seem to have different distributions. For example, the 1-nearest neighbor algorithm makes 5.7% on the USPS test set but only 2.3% on a test set of the same size randomly chosen from the full data set. (See, e.g., Freund and Schapire 1996.)

The USPS data set has been used in hundreds of papers and books; see, e.g., LeCun et al. 1990 and Vapnik 1998. Among the results reported in the literature for the error rate on the test set are: 2.5% for humans, 5.1% for the five-layer neural network LeNet 1, 4.0% for a polynomial support vector machine (with the polynomial kernel of degree 3). A very good result of 2.7% was obtained by Simard et al. 1993 without using any learning methods. Since we are usually interested in results for a randomly permuted USPS data set, the error rates that we obtain are not comparable to the error rates on the test set alone.

B.2 Boston Housing data set

The Boston Housing data set is a popular data set for testing different regression methods; it is available from several data repositories, such as Delve, the UCI data repository, and StatLib. The data set consists of 506 examples representing different areas of Boston, MA; each example has 12 continuous attributes, one binary attribute, and the label, which is the median house price in the area. This is the full list of all variables (13 attributes and the label):

- per capita crime rate;
- proportion of residential land zoned for lots over 25,000 square feet;
- proportion of non-retail business acres;
- binary variable indicating if the location is contiguous to the Charles River;
- nitric oxides concentration;
- average number of rooms per dwelling;
- proportion of owner-occupied units built prior to 1940;
- weighted distances to five Boston employment centers;
- index of accessibility to radial highways;
- full-value property-tax rate per \$10,000;
- pupil-teacher ratio;
- black population proportion;
- lower status population proportion;
- the label is the median value of owner-occupied homes.

All labels lie between \$5000 and \$50,000 and are given in units of \$1000. Similarly to the USPS data set, in our experiments we randomly permute the Boston Housing data set.

The following results for the mean squared difference between predicted and actual labels for the examples in a randomly chosen test set are reported in the literature: 12.4 and 11.7 for bagging, 10.7 for boosting, 7.7 and 7.2 for support vector machine.

B.3 Normalization

The performance of many machine-learning algorithms improves greatly if the objects are pre-processed. Suppose we are given examples

$$(x_1, y_1), \dots, (x_{n-1}, y_{n-1}),$$

a new object x_n , and our goal is to predict the label y_n ; each object x_i is a vector in \mathbb{R}^K and its components are denoted $x_{i,1}, \dots, x_{i,K}$. (In the case of the USPS data set, $K = 256$, and in the case of the Boston Housing data set, $K = 13$.)

The attributes of the Boston Housing data set have different orders of magnitude simply because of their nature: e.g., one attribute is binary and

another attribute is a distance; moreover, it is clear that the order of magnitude of different attributes often depends on the arbitrarily chosen unit of measurement. Therefore, it often leads to better results if we somehow normalize the columns of the matrix $x_{i,k}$. In our experiments with the Boston Housing data set we apply the linear transformation

$$x_{i,k} \mapsto x'_{i,k} := a_k + b_k x_{i,k}$$

such that, for all $k = 1, \dots, K$,

$$\min_i x'_{i,k} = -1, \quad \max_i x'_{i,k} = 1,$$

unless $\min_i x_{i,k} = \max_i x_{i,k}$, in which case we set $x'_{i,k} := 0$ for all i . As our experiments are on-line, we recompute a_k and b_k as each new object x_n arrives.

The kind of normalization required for the USPS data set is different: the attributes, being the intensities of individual pixels, are directly comparable, but the brightness and contrast of different hand-written images seems to be irrelevant to their classification. Therefore, it appears useful to normalize the rows of the matrix $x_{i,k}$. In many of our experiments we apply the linear transformation

$$x_{i,k} \mapsto x'_{i,k} := a_i + b_i x_{i,k},$$

where a_i and b_i are chosen such that, for all $i = 1, \dots, n$,

$$\frac{1}{K} \sum_{k=1}^K x'_{i,k} = 0, \quad \frac{1}{K} \sum_{k=1}^K (x'_{i,k})^2 = 1;$$

in practice, we never have the problem that $\min_k x_{i,k} = \max_k x_{i,k}$ (there are no absolutely uniform images). Since the pre-processing of each example is done independently of the other examples, even in on-line experiments we can do all pre-processing in advance.

No pre-processing is done in Chap. 7 on testing: all experiments are run on the original USPS data set; in the experiments reported in the other chapters this data set is randomly permuted and all the objects are normalized.

B.4 Randomization and reshuffling

Most of our experimental results involve randomization: our algorithms may be randomized (e.g., smooth conformal predictors are), or we might “reshuffle” a data set to make sure it conforms to an interesting assumption, such as exchangeability (when “reshuffling” is done, this is always mentioned explicitly). The main features of the graphs in this book are not significantly affected by the details of randomization.

All reported results are obtained by setting the initial state of MATLAB’s generator of pseudorandom numbers to 0, unless stated otherwise; since the

main purpose of computational experiments in this book is to illustrate theoretical results, we rarely use other initial states.

The most basic case of reshuffling is a random permutation of the data set performed to make sure that the assumption of exchangeability is satisfied. We do it for both USPS and Boston Housing data sets.

We have been saying “reshuffle” because the examples in the original data set are already in a somewhat random order, but Reality’s attempt at shuffling is often only half-hearted (see §7.1 for some results about the USPS data set), and we will sometimes say “shuffle”, especially when we are trying to achieve conformity with an assumption that is stronger than exchangeability.

After the general notion of an on-line compression model is introduced in §8.1, the idea of shuffling also becomes more general: given a data set z_1, \dots, z_n , we can find its summary $\sigma_n := t_n(z_1, \dots, z_n)$ and then draw another data sequence z'_1, \dots, z'_n from the conditional distribution $P_n(dz_1, \dots, dz_n | \sigma_n)$. If the model is very specific, shuffling is much more intrusive than a random permutation is, and might be better described as generation of a new data set sharing some characteristics with the original data set. (Shuffling for the Gauss linear model, described on p. 204, is of this type.)

In §9.4 we explicitly describe an efficient procedure of shuffling for junction-tree models, an important class of on-line compression models.

In all cases where shuffling is done in this book, Theorem 8.2 guarantees that at each significance level each smoothed conformal predictor makes errors independently at different trials with probability equal to the chosen significance level.

B.5 Bibliographical remarks

In our description of the USPS data set we partly followed LeCun et al. 1990. The original paper about the Boston Housing data set is Harrison and Rubinfeld 1978.

Most of the experimental results reproduced here are from Vapnik 1998. The average squared error of 7.2 (originally reported in Drucker et al. 1996) was achieved by the SVM using a polynomial kernel, with the degree chosen based on the performance on a validation set; Stitson et al. (1997) obtained a slightly weaker result 8.1 using a similar method, but this might be due to a random choice of the test set. The SVM using ANOVA splines achieves the average square error of 7.7 (Saunders et al. 1998). The average squared error of 11.7 for bagging is reported by Breiman (1994).

Appendix C: FAQ

In this short appendix we give our answers to several questions we have been asked by our colleagues and students. For simplicity we will discuss only prediction under unconstrained randomness, unless a different model is explicitly mentioned.

C.1 Unusual features of conformal prediction

Isn't your Proposition 2.4 (p. 27) too strong to be true? It is generally believed that to make categorical assertions about error probabilities some Bayes-type assumptions are needed and that the assumption of unconstrained randomness is not sufficient. For example, in the theory of PAC learning an error probability ϵ is only asserted with some probability $1 - \delta$.

It should be remembered that Proposition 2.4 does not assert that the probability of error, $\text{err}_n = 1$, is ϵ conditionally on knowing the whole past (2.4) (p. 19); it is only asserted that it is ϵ unconditionally and conditionally on knowing $\text{err}_1, \dots, \text{err}_{n-1}$. (Actually, it is quite obvious that the probability of error is often not equal to ϵ if the whole past is known: if the prediction set is empty, the conditional probability of error is 1; to balance this, the conditional probability that a non-empty prediction set is wrong will tend to be less than ϵ .)

How is it possible to achieve probability of error exactly ϵ in the problem of classification? For example, in the binary case there are only two possible labels, and you cannot expect that one of these labels will have probability exactly ϵ .

It is impossible to achieve the conditional probability of error equal to ϵ given the observed examples, but it is the unconditional probability of error that equals ϵ . Therefore, it implicitly involves averaging over different data sequences, and this gives us the leeway needed to obtain a probability precisely equal to ϵ .

Suppose the prediction set is empty at the chosen significance level. Does it mean that the result of conformal prediction is useless in this case?

Even if you are interested in only one significance level, say ϵ , the empty prediction still carries some information: you know that the object whose label you are predicting is unusual (in the long run the frequency of seeing such unusual objects is at most ϵ). Are you sure there was no mistake in recording the object? Do you still believe in the exchangeability assumption? If the answer to these questions is “yes” and you would still like to have a nonempty prediction, you have no choice but to look at what happens at the other significance levels. (As clear from Chap. 1, we share the standard view that it is never wise to concentrate on just one significance level.) Look at the smallest significance level ϵ'' at which the prediction set is empty and at the smallest significance level ϵ' at which the prediction set is not multiple. (Cf. the definition of confidence and credibility on p. 96.) If ϵ' is small and the difference between ϵ' and ϵ'' is significant, you can be fairly sure that the singular prediction set at the significance level $(\epsilon' + \epsilon'')/2$ will be correct.

C.2 Conformal prediction vs. standard methods

In your approach to classification, you have two main performance measures, Err_n (the number of errors) and Mult_n (the number of multiple predictions). You prove that Err_n grows as $n\epsilon$ (where ϵ is your chosen significance level) plus random noise and observe that in experiments Mult_n is usually small. In the standard approach (e.g., in the PAC theory) one trivially has $\text{Mult}_n = 0$ and observes that in experiments Err_n is usually small. There is a complete symmetry and you cannot claim that your approach is better.

This symmetry is superficial. Imagine that we are given a new object x_n having observed examples (x_i, y_i) , $i = 1, \dots, n - 1$. Suppose that for a small significance level ϵ a conformal predictor outputs a one-element prediction set $\{y\}$ and the standard approach outputs the simple prediction y . We can see that the prediction set $\{y\}$ is singular and we know that it has a small (equal to ϵ) probability of error. This gives us much more information than the simple prediction y does: in the latter case, we can see that y is singular but we knew in advance it was going to be singular; no reliable inference about the probability of error can be drawn from the smallness of the number of errors so far. (To use the language of the theory of martingales, the main asymmetry between Err_n and Mult_n is that Mult_n is predictable whereas Err_n is not.)

Could you summarize the main differences between conformal prediction and the standard approach to prediction?

Conformal predictors implement transductive rather than inductive learning. The basic validity result about conformal predictors is proved in the on-line

Table C.1. Three dichotomies for hedged prediction

inductive off-line statistical modeling	transductive on-line on-line compression modeling
---	---

rather than off-line learning protocol. To state our results in the simplest and most general form we use on-line compression rather than traditional statistical modeling. Table C.1 shows these three differences; conformal prediction is mostly concerned with the right-hand column and traditional machine learning is mostly concerned with the left-hand column.

Suppose I have a plausible on-line compression model M but know little about the set \mathcal{P} of probability distributions on \mathbf{Z}^∞ that agree with M ; in particular, \mathcal{P} may turn out to be empty or a singleton. Should I be worried about this?

In our opinion, in practical applications you can safely ignore the foundational questions such as whether \mathcal{P} is rich enough. You know that for each finite horizon N there are plenty of probability distributions on \mathbf{Z}^N that agree with M , and you are never going to reach infinity.

In Chaps. 2–4 you show how one can use standard machine-learning methods to devise nonconformity measures. Are there any formal connections between those standard methods and conformal predictors based on those methods?

We are not aware of any formal connections that always hold; it is often true, however, that the simple prediction produced by a machine-learning method will belong to the prediction set produced by the corresponding conformal predictor, unless that prediction set is empty.

Notation and abbreviations

Sets, bags, and sequences

\emptyset	the empty set
\mathbb{N}	the positive integer numbers, $\{1, 2, \dots\}$
\mathbb{N}_0	the nonnegative integer numbers, $\{0, 1, \dots\}$
\mathbb{Z}	the integer numbers
\mathbb{Q}	the rational numbers
\mathbb{R}	the real numbers
$\overline{\mathbb{R}}$	the extended real numbers, $\mathbb{R} \cup \{-\infty, \infty\}$
$\{z_1, \dots, z_n\}$	set (each element enters only once)
$\{z_1, \dots, z_n\}$	bag (can contain more than one copy of the same element; p. 23)
(z_1, \dots, z_n)	sequence (the parentheses and commas may be omitted)
\square	the empty sequence
$[a_1, \dots, a_n]$	the set of all infinite continuations of a finite sequence a_1, \dots, a_n
$ A $	the size of a set or bag A
Z^n	the set of all sequences of elements of Z of length n
$Z^{(n)}$	the set of all bags of elements of Z of size n
Z^*	the set of all finite sequences of elements of Z
$Z^{(*)}$	the set of all bags (always finite) of elements of Z
Z^∞	the set of all infinite sequences of elements of Z
2^Z	the set of all subsets of a set Z
Y^X	the set of all functions of the type $X \rightarrow Y$
$A(z)$	the element containing $z \in Z$ of a partition A of a set Z

Stochastics

\mathbb{P}	probability
\mathbb{E}	expectation
$\mathbf{P}(Z)$	the set of all probability distributions on Z (measurable space)
\mathbf{B}_δ	the Bernoulli distribution on $\{0, 1\}$ with the parameter δ : $\mathbf{B}_\delta\{1\} = \delta$ and $\mathbf{B}_\delta\{0\} = 1 - \delta$
$\mathbf{N}_{\mu, \sigma^2}$	the normal distribution on \mathbb{R} with mean μ and variance σ^2
\mathbf{U}	the uniform distribution on $[0, 1]$
$t_{\delta, n}$	the percentage point of the t -distribution: $\mathbb{P}\{\xi \geq t_{\delta, n}\} = \delta$, where ξ has Student's t -distribution with n degrees of freedom
\mathbf{z}_δ	the percentage point of the standard normal distribution: $\mathbb{P}\{\xi \geq \mathbf{z}_\delta\} = \delta$, where ξ has the normal distribution $\mathbf{N}_{0, 1}$
$Q_{\mathbf{X}}$	the marginal distribution of $Q \in \mathbf{P}(\mathbf{X} \times \mathbf{Y})$ on \mathbf{X} (p. 65)
$Q_{\mathbf{Y} \mathbf{X}}$	the regular conditional distribution of $y \in \mathbf{Y}$ given $x \in \mathbf{X}$, where (x, y) is distributed as Q (p. 65)
$\Omega \hookrightarrow Z$	Markov kernel from Ω to Z (p. 280)
Pf^{-1}	the image of P under a mapping f (p. 276)
ρ	variation distance between probability distributions (p. 163)
Π	the game space (p. 148)

Machine learning

\mathbf{X}	object space (p. 17)
\mathbf{Y}	label space, $ \mathbf{Y} > 1$ (p. 17)
\mathbf{Z}	the example space ($\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, p. 18)
$\tilde{\mathbf{X}}$	the extended object space $\mathbf{X} \times [0, 1]$ (p. 61)
$\tilde{\mathbf{Z}}$	the extended example space $\mathbf{X} \times [0, 1] \times \mathbf{Y}$ (p. 61)
\mathbf{H}	feature space (with the feature mapping $F : \mathbf{X} \rightarrow \mathbf{H}$; p. 36)

Confidence prediction

ϵ	significance level
Γ_n^ϵ	the prediction set at trial n (p. 20)
err_n^ϵ	the indicator of error at trial n (p. 20)
$\overline{\text{err}}$	the predictable version of err (p. 77)
Err_n^ϵ	the cumulative number of errors up to trial n (p. 20)

$\overline{\text{Err}}$	the predictable version of Err (p. 78)
mult_n^ϵ	the indicator of multiple prediction at trial n (p. 54)
$\overline{\text{mult}}$	the predictable version of mult (p. 78)
Mult_n^ϵ	the cumulative number of multiple predictions up to trial n (p. 54)
$\overline{\text{Mult}}$	the predictable version of Mult (p. 78)
emp_n^ϵ	the indicator of empty prediction at trial n (p. 54)
$\overline{\text{emp}}$	the predictable version of emp (p. 80)
Emp_n^ϵ	the cumulative number of empty predictions up to trial n (p. 54)
$\overline{\text{Emp}}$	the predictable version of Emp (p. 80)
τ_n	the n th random number used by a randomized confidence predictor (p. 22)
τ'_n, τ''_n	the two components of τ_n , as defined in §3.3 (p. 61)
$\#\sigma$	the number of data sequences generating a summary σ (pp. 209, 226)
$f(x)$	the predictability of object $x \in \mathbf{X}$ (p. 65)
F	the predictability distribution function (p. 65)
M	the multiplicity curve (p. 66)
E	the emptiness curve (p. 66)
ϵ_0	the critical significance level (p. 68)

Other notation

\mathbb{I}_A	the indicator function of a set or property A (p. 59)
$f _A$	the restriction of a function or kernel f to a subset A of its domain
$\text{diam } A$	the diameter (largest distance between points) of A
$\text{co } A$	the convex hull of a set A in a linear space
I_n	the identity $n \times n$ matrix (n is omitted if clear from the context)
X'	matrix X transposed
X^{-1}	the inverse of matrix X
$\text{rank } X$	the rank of matrix X
$u \vee v$	the maximum of u and v , also denoted $\max(u, v)$
$u \wedge v$	the minimum of u and v , also denoted $\min(u, v)$
u^+	$u \vee 0$
u^-	$(-u) \vee 0$

$F(t-)$	the limit of $F(u)$ as u approaches t from below
$F(t+)$	the limit of $F(u)$ as u approaches t from above
$\forall^\infty n$	from some n on
$f_n = O(g_n)$	$\limsup_{n \rightarrow \infty} (f_n/g_n) < \infty$ (used for $f_n, g_n > 0$)
$f_n = \Theta(g_n)$	$f_n = O(g_n)$ and $g_n = O(f_n)$

Abbreviations

FCVP	fully conditional Venn predictor
ICP	inductive conformal predictor
LSCM	least squares confidence machine
MCP	Mondrian conformal predictor
MCT	Mondrian conformal transducer
NNR	nearest neighbors regression
OCM	on-line compression model
RRCM	ridge regression confidence machine
SVM	support vector machine
USPS	see §B.1 (p. 291)

References

- Martin ANTHONY (2003). Boolean functions and artificial neural networks. CDAM Research Report LSE-CDAM-2003-01, Centre for Discrete and Applicable Mathematics, London School of Economics and Political Science.
- Martin ANTHONY and Peter L. BARTLETT (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, England.
- John ARBUTHNOTT (1710–1712). An argument for divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27:186–190.
- Raymond C. ARCHIBALD (1926). A rare pamphlet of Moivre and some of his discoveries. *Isis*, 8:671–683. A facsimile copy of De Moivre's 1733 pamphlet starts on p. 677.
- Nachman ARONSAJN (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404.
- Eugene A. ASARIN (1987). Some properties of Kolmogorov δ -random finite sequences. *Theory of Probability and its Applications*, 32:507–508.
- Eugene A. ASARIN (1988). On some properties of finite objects random in the algorithmic sense. *Soviet Mathematics Doklady*, 36:109–112. The Russian original published in 1987.
- Kazuoki AZUMA (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367.
- George A. BAKER (1935). The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample. *Annals of Mathematical Statistics*, 6: 197–201.
- Tadeusz BANACHIEWICZ (1937a). Sur l'inverse d'un cracovien et une solution générale d'un système d'équations linéaires. *Comptes rendus mensuels des Séances de la Classe des Sciences Mathématiques et Naturelles de l'Académie Polonaise des Sciences et des Lettres*, 4:3–4.
- Tadeusz BANACHIEWICZ (1937b). Zur Berechnung der Determinanten, wie auch der Inversen, und zur darauf basierten Auflösung der Systeme linearer Gleichungen. *Acta Astronomica, Series C*, 3:41–67.
- George A. BARNARD (1977). Pivotal inference and the Bayesian controversy. *Bulletin of the International Statistical Institute*, 47:543–551.

- Thomas BAYES (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.
- James O. BERGER and Mohan DELAMPADY (1987). Testing precise hypotheses (with discussion). *Statistical Science*, 2:317–352.
- José M. BERNARDO and Adrian F. M. SMITH (1994). *Bayesian Theory*. Wiley, Chichester, England. Paperback edition: 2000.
- Jacob BERNOULLI (1713). *Ars Conjectandi*. Thurnisius, Basel. Russian translation (second edition, with commentaries by Oscar B. Sheynin and Yurii V. Prokhorov): О законе больших чисел, Nauka, Moscow, 1986.
- Christopher M. BISHOP (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, England.
- Nicolas BOURBAKI (1958). *Éléments de mathématique, Livre IV, Fonctions d'une variable réelle (théorie élémentaire)*. Hermann, Paris, second edition.
- Nicolas BOURBAKI (1969). *Éléments de mathématique, Livre VI, Intégration, Chapitre IX, Intégration sur les espaces topologiques séparé, Note historique*. Hermann, Paris.
- Leo BREIMAN (1994). Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkley.
- Lawrence D. BROWN, T. Tony CAI, and Anirban DASGUPTA (2001). Interval estimation for a binomial proportion (with discussion). *Statistical Science*, 16: 101–133.
- Francesco P. CANTELLI (1933). Sulla determinazione empirica della leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:421–424.
- Aleksei Ya. CHERVONENKIS (2004). Воспоминания Червоненкиса (Chervonenkis's recollections). Manuscript.
- R. Dennis COOK and Sanford WEISBERG (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.
- Thomas H. CORMEN, Charles E. LEISERSON, Ronald L. RIVEST, and Clifford STEIN (2001). *Introduction to Algorithms*. MIT Press, Cambridge, MA, second edition.
- Robert G. COWELL, A. Philip DAWID, Steffen L. LAURITZEN, and David J. SPIEGELHALTER (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.
- David R. COX (1958a). The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society, Series B*, 20:215–242.
- David R. COX (1958b). Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372.
- David R. COX (1970). *The Analysis of Binary Data*. Methuen, London. Second edition: 1989 (with E. J. Snell).
- David R. COX and David V. HINKLEY (1974). *Theoretical Statistics*. Chapman and Hall, London.
- Harald CRAMÉR (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.
- Noel A. C. CRESSIE (1993). *Statistics for Spatial Data*. Wiley, New York, revised edition.
- Nello CRISTIANINI and John SHawe-TAYLOR (2000). *An Introduction to Support Vector Machines and Other Kernel-based Methods*. Cambridge University Press, Cambridge, England.
- Anthony C. DAVISON and David V. HINKLEY (1997). *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge, England.

- A. Philip DAWID (1982). Intersubjective statistical models. In George S. Koch and Fabio Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 217–232. North-Holland, Amsterdam.
- A. Philip DAWID (1985). Self-calibrating priors do not exist: Comment. *Journal of the American Statistical Association*, 80:340–341. This is a contribution to the discussion in Oakes (1985).
- A. Philip DAWID (1986). Probability forecasting. In Samuel Kotz, Norman L. Johnson, and Campbell B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley, New York.
- A. Philip DAWID and Mervyn STONE (1982). The functional-model basis of fiducial inference (with discussion). *Annals of Statistics*, 10:1054–1067.
- Abraham DE MOIVRE (1733). Approximatio ad summam terminorum binomii $a + b$ ⁿ in seriem expansi. Included in Archibald 1926.
- Morris H. DEGROOT (1988). A conversation with George A. Barnard. *Statistical Science*, 3:196–212.
- Philip DERBEKO, Ran EL-YANIV, and Ron MEIR (2004). Error bounds for transductive learning via compression and clustering. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA. The journal version: Explicit learning curves for transduction and application to clustering and compression algorithms; to appear in *Journal of Artificial Intelligence Research*.
- Luc DEVROYE, László GYÖRFI, Adam KRZYŻAK, and Gábor LUGOSI (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385.
- Luc DEVROYE, László GYÖRFI, and Gábor LUGOSI (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- Luc DEVROYE and Gábor LUGOSI (2001). *Combinatorial Methods in Density Estimation*. Springer, New York.
- Persi DIACONIS and David A. FREEDMAN (1980). De Finetti's theorem for Markov chains. *Annals of Probability*, 8:115–130.
- Thomas G. DIETTERICH and Ghulum BAKIRI (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286.
- Jean DIEUDONNÉ (1948). Sur le théorème de Lebesgue–Nikodym. III. *Annales de l'Institut Fourier*, 23:25–53.
- Joseph L. DOOB (1953). *Stochastic Processes*. Wiley, New York.
- Norman R. DRAPER and Harry SMITH (1998). *Applied Regression Analysis*. Wiley, New York, third edition.
- Harris DRUCKER, Chris J. C. BURGES, Linda KAUFMAN, Alex SMOLA, and Vladimir N. VAPNIK (1996). Support Vector regression machines. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, Cambridge, MA.
- Richard M. DUDLEY (2002). *Real Analysis and Probability*. Cambridge University Press, Cambridge, England. Original edition published in 1989 by Wadsworth.
- Aryeh DVORETZKY, Jack C. KIEFER, and Jacob WOLFOWITZ (1956). Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669.
- Bradley EFRON (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

- Bradley EFRON (2003). Second thoughts on the bootstrap. *Statistical Science*, 18: 135–140.
- Bradley EFRON and Robert J. TIBSHIRANI (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Bruno de FINETTI (1930). Funzione caratteristica di un fenomeno aleatorio. *Memorie della Reale Accademia Nazionale dei Lincei: Classe di scienze fisiche, matematiche, e naturali, Serie 6*, 4:86–133.
- Bruno de FINETTI (1937). La prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68. An English translation of this article is included in Kyburg and Smokler 1980.
- Bruno de FINETTI (1938). *Sur la condition d'équivalence partielle*, volume 739 of *Actualités Scientifiques et Industrielles*. Hermann, Paris. An English translation is included in Jeffrey 1981, pp. 193–206.
- Ronald A. FISHER (1922). The goodness of fit of regression formulae and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85: 597–612.
- Ronald A. FISHER (1925). Applications of “Student’s” distribution. *Metron*, 5: 90–104.
- Ronald A. FISHER (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.
- Ronald A. FISHER (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398.
- Ronald A. FISHER (1973a). *Statistical Methods and Scientific Inference*. Hafner, New York, third edition. Included in Fisher 1991. First edition: 1956.
- Ronald A. FISHER (1973b). *Statistical Methods for Research Workers*. Hafner, New York, fourteenth (revised and enlarged) edition. Included in Fisher 1991. First edition: 1925.
- Ronald A. FISHER (1991). *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, Oxford, England. Edited by J. Henry BENNETT.
- Sally FLOYD and Manfred K. WARMUTH (1995). Sample compression, learnability, and the Vapnik–Chervonenkis dimension. *Machine Learning*, 21:269–304.
- Donald A. S. FRASER (1951). Sequentially determined statistically equivalent blocks. *Annals of Mathematical Statistics*, 22:372–381.
- Donald A. S. FRASER (1953). Nonparametric tolerance regions. *Annals of Mathematical Statistics*, 24:44–55.
- Donald A. S. FRASER (1957). *Nonparametric Methods in Statistics*. Wiley, New York.
- Donald A. S. FRASER and Irwin GUTTMAN (1956). Tolerance regions. *Annals of Mathematical Statistics*, 27:16–32.
- David A. FREEDMAN (1962). Invariants under mixing which generalise de Finetti’s theorem. *Annals of Mathematical Statistics*, 33:916–923.
- David A. FREEDMAN (1963). Invariants under mixing which generalise de Finetti’s theorem: Continuous time parameter. *Annals of Mathematical Statistics*, 34: 1194–1216.
- Yoav FREUND, Yishay MANSOUR, and Robert E. SCHAPIRE (2004). Generalization bounds for averaged classifiers. *Annals of Statistics*, 32:1698–1722.

- Yoav FREUND and Robert E. SCHAPIRE (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156.
- Yoav FREUND and Robert E. SCHAPIRE (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139. Conference version: EuroCOLT'1995.
- Peter GÁCS (1980). Exact expressions for some randomness tests. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 26:385–394.
- Alex GAMMERMAN (1997). Machine learning: Progress and prospects. Royal Holloway, University of London, Egham, Surrey, TW20 0EX, England. This booklet is based on an inaugural lecture delivered on 11 December 1996.
- Alex GAMMERMAN, Vladimir VAPNIK, and Vladimir VOVK (1997). Transduction in pattern recognition. Manuscript submitted to the Fifteenth International Joint Conference on Artificial Intelligence in January 1997. Extended version published as Gammerman et al. 1998. The algorithm proposed in this paper was described in a 1996 public lecture (Gammerman 1997).
- Alex GAMMERMAN, Vladimir Vovk, and Vladimir VAPNIK (1998). Learning by transduction. In Gregory F. Cooper and Serafín Moral, editors, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–155, San Francisco, CA. Morgan Kaufmann.
- Carl F. GAUSS (1823). *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Pars Posterior*. Dieterich, Göttingen.
- Valery I. GLIVENKO (1933). Sulla determinazione empirica di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:92–99.
- Irwin GUTTMAN (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Griffin, London.
- Jules HAAG (1928). Sur une problème général de probabilités et ses diverses applications. In *Proceedings of the International Congress of Mathematicians, Toronto, 1924*, pages 659–674, Toronto. Toronto University Press.
- David HARRISON and Daniel L. RUBINFELD (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.
- David HAUSSLER, Nick LITTLESTONE, and Manfred K. WARMUTH (1990). Predicting {0, 1}-functions on randomly drawn points. Technical Report UCSC-CRL-90-54, Computer Research Laboratory, University of California, Santa Cruz. Lemma 2.6 and Theorem 2.2 of this technical report are not stated explicitly in its final version Haussler et al. (1994).
- David HAUSSLER, Nick LITTLESTONE, and Manfred K. WARMUTH (1994). Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115:248–292. Preliminary versions in: Proceedings of the Twenty-Ninth Annual Symposium on Foundations of Computer Science, pp. 100–109 (1988); Proceedings of the First Workshop on Computational Learning Theory, pp. 280–296 (1988).
- Harold V. HENDERSON and Shayle R. SEARLE (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23:53–60.
- Ralf HERBRICH and Robert C. WILLIAMSON (2002). Learning and generalization: Theoretical bounds. In Michael A. Arbib, editor, *Handbook of Brain Theory and Neural Networks*, pages 3140–3150. MIT Press, Cambridge, MA, second edition.
- Mark HERBSTER and Manfred K. WARMUTH (1998). Tracking the best expert. *Machine Learning*, 32:151–178.

- Edwin HEWITT and Leonard J. SAVAGE (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501.
- David C. HOAGLIN and Roy E. WELSCH (1978). The hat matrix in regression and ANOVA. *American Statistician*, 32:17–22.
- Wassily HOEFFDING (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.
- Arthur E. HOERL (1959). Optimum solution of many variables equations. *Chemical Engineering Progress*, 55:69–78.
- Arthur E. HOERL (1962). Applications of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59.
- Arthur E. HOERL and Robert W. KENNARD (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.
- Arthur E. HOERL and Robert W. KENNARD (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82. Erratum, 12:723.
- Roger W. HOERL (1985). Ridge analysis 25 years later. *American Statistician*, 39: 186–192.
- David W. HOSMER and Stanley LEMESHOW (2000). *Applied Logistic Regression*. Wiley, New York, second edition.
- David HUME (1739–1740). *A Treatise of Human Nature*. Noon (vols. 1–2, 1739) and Longman (vol. 3, 1740), London.
- Richard C. JEFFREY, editor (1981). *Studies in Inductive Logic and Probability*. University of California Press, Berkley, CA.
- Finn V. JENSEN (1996). *An Introduction to Bayesian Networks*. UCL Press, London.
- Marek KARPINSKI and Angus J. MACINTYRE (1995). Polynomial bounds for VC dimension of sigmoidal neural networks. In *Proceeding of the Twenty-Seventh Annual ACM Symposium on the Theory of Computing*, pages 200–208, New York. ACM Press.
- Marek KARPINSKI and Angus J. MACINTYRE (1997). Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, 54:169–176.
- Johan H. B. KEMPERMAN (1956). Generalized tolerance limits. *Annals of Mathematical Statistics*, 27:180–186.
- Aleksandr Ya. KHINCHIN (1932). Sur les classes d'événements équivalents. *Математический сборник*, 39:40–43.
- Berna E. KILINÇ (2001). The reception of John Venn's philosophy of probability. In Vincent F. Hendricks, Stig Andur Pedersen, and Klaus Frovin Jørgensen, editors, *Probability Theory: Philosophy, Recent History and Relations to Science*, pages 97–121. Kluwer, Dordrecht.
- John F. C. KINGMAN (1972). On random sequences with spherical symmetry. *Biometrika*, 59:492–493.
- John F. C. KINGMAN (1978). Uses of exchangeability. *Annals of Probability*, 6: 183–197.
- Andrei N. KOLMOGOROV (1928). Sur une formule limite de M. A. Khintchine. *Comptes rendus des Séances de l'Académie des Sciences*, 186:824–825.
- Andrei N. KOLMOGOROV (1929). Über das Gesetz des iterierten Logarithmus. *Mathematische Annalen*, 101:126–135.
- Andrei N. KOLMOGOROV (1933a). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. Published in English as *Foundations of the Theory of Probabil-*

- ity* (Chelsea, New York). Translation edited by Nathan Morrison. First edition, 1950; second edition, 1956.
- Andrei N. KOLMOGOROV (1933b). Sulla determinazione empirica di unna legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.
- Andrei N. KOLMOGOROV (1942). Определение центра рассеивания и меры точности по ограниченному числу наблюдений (The estimation of the mean and precision from a finite sample of observations). *Известия АН СССР, серия математическая*, 6:3–32.
- Andrei N. KOLMOGOROV (1963). On tables of random numbers. *Sankhya, The Indian Journal of Statistics, Series A*, 25:369–376.
- Andrei N. KOLMOGOROV (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7.
- Andrei N. KOLMOGOROV (1968). Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664.
- Andrei N. KOLMOGOROV (1983). Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40. This article was written in 1970 in connection with Kolmogorov's talk at the International Mathematical Congress in Nice.
- Henry KYBURG and Henry SMOKLER, editors (1980). *Studies in Subjective Probability*. Krieger, New York, second edition.
- John LANGFORD (2004). Tutorial on practical prediction theory for classification. Article based on an ICML'2003 tutorial and available from the author's web site. Accessed on March 28, 2004.
- John LANGFORD and John SHawe-TAYLOR (2003). PAC-Bayes and margins. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA.
- Pierre Simon LAPLACE (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de mathématique et de physique, présentés à l'Académie royale des sciences, par divers savans & lus dans ses assemblées*, 6:621–656. English translation: *Statistical Science*, 1:364–378.
- Steffen L. LAURITZEN (1982). *Statistical Models as Extremal Families*. Aalborg University Press, Aalborg, Denmark.
- Steffen L. LAURITZEN (1988). *Extremal Families and Systems of Sufficient Statistics*, volume 49 of *Lecture Notes in Statistics*. Springer, New York.
- Steffen L. LAURITZEN (1996). *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. Clarendon Press, Oxford, England.
- Yann LECUN, Bernhard E. BOSER, John S. DENKER, Donnie HENDERSON, R. E. HOWARD, Wayne E. HUBBARD, and Lawrence D. JACKEL (1990). Handwritten digit recognition with a back-propagation network. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan Kaufmann, San Mateo, CA. The surname of the first author is spelt as Le Cun in this paper; we use the spelling that Prof. LeCun now prefers.
- Adrien M. LEGENDRE (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Courcier, Paris.
- Erich L. LEHMANN (1986). *Testing Statistical Hypotheses*. Wiley, New York, second edition. First edition: 1959.
- Leonid A. LEVIN (1973). On the notion of a random sequence. *Soviet Mathematics Doklady*, 14:1413.

- Leonid A. LEVIN (1976). Uniform tests of randomness. *Soviet Mathematics Doklady*, 17:337–340.
- Leonid A. LEVIN (1984). Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61:15–37.
- Paul LÉVY (1922). *Leçons d'Analyse fonctionnelle*. Gauthier-Villars, Paris.
- Ming LI and Paul VITÁNYI (1997). *An Introduction to Kolmogorov Complexity and its Applications*. Springer, New York, second edition.
- Yi LI, Philip M. LONG, and Aravind SRINIVASAN (2001). The one-inclusion graph algorithm is near-optimal for the prediction model of learning. *IEEE Transactions on Information Theory*, 47:1257–1261.
- Nick LITTLESTONE and Manfred K. WARMUTH (1986). Relating data compression and learnability. Technical report, University of California, Santa Cruz.
- Andrei A. MARKOV (1906). Распространение закона больших чисел на величины, зависящие друг от друга (Extension of the law of large numbers to dependent events). *Notices of the Physico-Mathematical Society of the University of Kazan, Second Series*, 15:135–156.
- Per MARTIN-LÖF (1966). The definition of random sequences. *Information and Control*, 9:602–619.
- Per MARTIN-LÖF (1974). Repetitive structures. In Ole E. Barndorff-Nielsen, Preben Blæsild, and Geert Schou, editors, *Proceedings of Conference on Foundational Questions in Statistical Inference*, Aarhus. Memoirs 1.
- Pascal MASSART (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Annals of Probability*, 18:1269–1283.
- Georges MATHERON (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.
- David A. MCALLESTER (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York. ACM Press. Journal version: McAllester 1999b.
- David A. MCALLESTER (1999a). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 164–170, New York. ACM Press.
- David A. MCALLESTER (1999b). Some PAC-Bayesian theorems. *Machine Learning*, 37:355–363.
- Colin J. H. McDIARMID (1989). On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics: Invited Papers at the Twelfth British Combinatorial Conference*, volume 141 of *London Mathematical Society Lecture Notes Series*, pages 148–188. Cambridge University Press, Cambridge, England.
- Thomas MELLUISH (2005). *Transductive algorithms for finding confidence information for regression estimation in the typicalness framework*. PhD thesis, Royal Holloway, University of London. Submitted.
- Thomas MELLUISH, Craig SAUNDERS, Ilia NOURETDINOV, and Vladimir VOVK (2001a). Comparing the Bayes and typicalness frameworks. Technical Report CLRC-TR-01-05, Computer Learning Research Centre, Royal Holloway, University of London.
- Thomas MELLUISH, Craig SAUNDERS, Ilia NOURETDINOV, and Vladimir VOVK (2001b). Comparing the Bayes and typicalness frameworks. In Luc De Raedt and Peter A. Flach, editors, *Machine Learning: ECML'2001. Proceedings of the Twelfth European Conference on Machine Learning*, volume 2167 of *Lecture*

- Notes in Computer Science*, pages 360–371, Heidelberg. Springer. The formula for the variance of the predictive distribution in terms of a kernel (§4 of this paper) is wrong; this is corrected in Melluish et al. 2001a and Melluish 2005.
- James MERCER (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, 209:415–446.
- Richard von MISES (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99.
- Richard von MISES (1928). *Wahrscheinlichkeitsrechnung, Statistik und Wahrheit*. Julius Springer, Wien. The second German edition appeared in 1936 and the third in 1951. English editions, under the title *Probability, Statistics and Truth*, appeared in 1939 and 1957.
- Melanie MITCHELL (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.
- Tom M. MITCHELL (1997). *Machine Learning*. McGraw-Hill, New York.
- Douglas C. MONTGOMERY, Elizabeth A. PECK, and G. Geoffrey VINING (2001). *Introduction to Linear Regression Analysis*. Wiley, New York, third edition.
- NEUROCOLT (2002). Generalisation bounds less than 0.5. NeuroCOLT Workshop, 29 April – 2 May 2002, Windsor, England.
- Ilia NOURETDINOV, Thomas MELLUSH, and Vladimir VOVK (2001a). Ridge Regression Confidence Machine. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 385–392, San Francisco, CA. Morgan Kaufmann.
- Ilia NOURETDINOV and Vladimir VOVK (2003). Criterion of calibration for transductive confidence machine with limited feedback. In Ricard Gavaldà, Klaus P. Jantke, and Eiji Takimoto, editors, *Proceedings of the Fourteenth International Conference on Algorithmic Learning Theory*, volume 2842 of *Lecture Notes in Artificial Intelligence*, pages 259–267, Berlin. Springer. To appear in *Theoretical Computer Science* (Special Issue devoted to the ALT’2003).
- Ilia NOURETDINOV, Vladimir VOVK, Michael VYUGIN, and Alex GAMMERMAN (2001b). Pattern recognition and density estimation under the general i.i.d. assumption. In David Helmbold and Bob Williamson, editors, *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory and Fifth European Conference on Computational Learning Theory*, volume 2111 of *Lecture Notes in Artificial Intelligence*, pages 337–353. Springer.
- Ilia NOURETDINOV, Vladimir V. YUGIN, and Alex GAMMERMAN (2003). Transductive confidence machine is universal. In Ricard Gavaldà, Klaus P. Jantke, and Eiji Takimoto, editors, *Proceedings of the Fourteenth International Conference on Algorithmic Learning Theory*, volume 2842 of *Lecture Notes in Artificial Intelligence*, pages 283–297, Berlin. Springer.
- David OAKES (1985). Self-calibrating priors do not exist (with discussion). *Journal of the American Statistical Association*, 80:339–342.
- Anthony O’HAGAN (1994). *Kendall’s Advanced Theory of Statistics*, volume 2b: Bayesian Inference. Arnold, London.
- Harris PAPADOPOULOS (2004). *Qualified predictions for large data sets*. PhD thesis, Royal Holloway, University of London.
- Harris PAPADOPOULOS, Konstantinos PROEDROU, Vladimir VOVK, and Alex GAMMERMAN (2002a). Inductive Confidence Machines for regression. In *Machine Learning: ECML’2002. Proceedings of the Thirteenth European Conference on*

- Machine Learning*, volume 2430 of *Lecture Notes in Computer Science*, pages 345–356.
- Harris PAPADOPOULOS, Vladimir VOVK, and Alex GAMMERMAN (2002b). Qualified predictions for large data sets in the case of pattern recognition. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA'02)*, pages 159–163. CSREA Press.
- Judea PEARL (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA.
- Egon S. PEARSON (1968). Studies in the history of probability and statistics. XX: Some early correspondence between W. S. Gosset, R. A. Fisher and Karl Pearson, with notes and comments. *Biometrika*, 55:445–457.
- Karl PEARSON (1925). James Bernoulli's theorem. *Biometrika*, 17:201–210.
- Richard R. PICARD and Kenneth N. BERK (1990). Data splitting. *American Statistician*, 44:140–147.
- Robin L. PLACKETT (1972). Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares. *Biometrika*, 59:239–251.
- Karl R. POPPER (1999). *The Logic of Scientific Discovery*. Routledge, London. First published in German in 1934; first English edition 1959.
- Konstantinos PROEDROU (2003). *Rigorous measures of confidence for pattern recognition and regression*. PhD thesis, Royal Holloway, University of London.
- Yurii V. PROKHOROV (1986). Закон больших чисел и оценки вероятностей больших уклонений. This is Commentary II to the second Russian edition of Jacob Bernoulli's *Ars conjectandi* (Bernoulli 1713).
- J. Ross QUINLAN (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
- Nancy REID (1994). A conversation with Sir David Cox. *Statistical Science*, 9: 439–455.
- Brian D. RIPLEY (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, England.
- Ronald L. RIVEST and Robert H. SLOAN (1988). Learning complicated concepts reliably and usefully. In *Proceedings of the First Annual Workshop on Computational Learning Theory*, pages 69–79, San Mateo, CA. Morgan Kaufmann.
- L. Chris G. ROGERS and David WILLIAMS (1994). *Diffusions, Markov Processes, and Martingales*, volume 1: Foundations. Wiley, Chichester, England, second edition. Reissued by Cambridge University Press in *Cambridge Mathematical Library*, 2000.
- David E. RUMELHART and James L. MCCLELLAND, editors (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations. MIT Press, Cambridge, MA.
- Daniil RYABKO (2003). Relaxing i.i.d. assumption in online pattern recognition. Technical Report CS-TR-03-11, Department of Computer Science, Royal Holloway, University of London.
- Daniil RYABKO, Vladimir VOVK, and Alex GAMMERMAN (2003). Online prediction with real teachers. Technical Report CS-TR-03-09, Department of Computer Science, Royal Holloway, University of London.
- Craig SAUNDERS (2000). *Efficient implementation and experimental testing of transductive algorithms for predicting with confidence*. PhD thesis, Royal Holloway, University of London.

- Craig SAUNDERS, Alex GAMMERMAN, and Vladimir VOVK (1998). Ridge regression learning algorithm in dual variables. In Jude W. Shavlik, editor, *Machine Learning, Proceedings of the Fifteenth International Conference*, pages 515–521, San Francisco, CA. Morgan Kaufmann.
- Craig SAUNDERS, Alex GAMMERMAN, and Vladimir VOVK (1999). Transduction with confidence and credibility. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 722–726. Morgan Kaufmann.
- Craig SAUNDERS, Alexander GAMMERMAN, and Vladimir Vovk (2000). Computationally efficient transductive machines. In Hiroki Arimura, Sanjay Jain, and Arun Sharma, editors, *Proceedings of the Eleventh International Conference on Algorithmic Learning Theory (ALT 2000)*, volume 1968 of *Lecture Notes in Artificial Intelligence*, pages 325–333, Berlin. Springer.
- Robert E. SCHAPIRE (1990). The strength of weak learnability. *Machine Learning*, 5:197–227.
- Robert E. SCHAPIRE, Yoav FREUND, Peter L. BARTLETT, and Wee Sun LEE (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686.
- Mark J. SCHERVISH (1995). *Theory of Statistics*. Springer, New York.
- Bernhard SCHÖLKOPF and Alexander J. SMOLA (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.
- Hilary L. SEAL (1967). Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model. *Biometrika*, 54:1–24.
- Matthias SEEGER (2003). PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269.
- Glenn SHAFER (1996a). *The Art of Causal Conjecture*. MIT Press, Cambridge, MA.
- Glenn SHAFER (1996b). *Probabilistic Expert Systems*, volume 67 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.
- Glenn SHAFER and Vladimir VOVK (2001). *Probability and Finance: It's only a Game!* Wiley, New York.
- Glenn SHAFER and Vladimir VOVK (2003). Sources of Kolmogorov's *Grundbegriffe*, The Game-Theoretic Probability and Finance Project, Working Paper #4, <http://probabilityandfinance.com>.
- John SHAWE-TAYLOR and Nello CRISTIANINI (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, England.
- Walter A. SHEWHART (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand, New York.
- Oscar B. SHEYNIN (2003). История теории вероятностей до XX века (*The History of Probability Theory before the Twentieth Century*). Северо-Западный заочный государственный технический университет, St. Petersburg.
- Albert N. SHIRYAEV (1996). *Probability*. Springer, New York, second edition.
- Patrice SIMARD, Yann LE CUN, and John DENKER (1993). Efficient pattern recognition using a new transformation distance. In Steven J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, San Mateo, CA.
- Adrian F. M. SMITH (1981). On random sequences with centred spherical symmetry. *Journal of the Royal Statistical Society, Series B*, 43:208–209.
- J. Laurie SNELL (1997). A conversation with Joe Doob. *Statistical Science*, 12: 301–311.

- Stephen M. STIGLER (1981). Gauss and the invention of least squares. *Annals of Statistics*, 9:465–474. Revised version published as Chap. 17 of Stigler 1999.
- Stephen M. STIGLER (1986a). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA.
- Stephen M. STIGLER (1986b). Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378.
- Stephen M. STIGLER (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA.
- Robert A. STINE (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80:1026–1031.
- Mark O. STITSON, Alex GAMMERMAN, Vladimir N. VAPNIK, Vladimir VOVK, Chris WATKINS, and Jason WESTON (1997). Support Vector ANOVA decomposition. Technical Report CSD-TR-97-22, Royal Holloway, University of London.
- Charles J. STONE (1977). Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5:595–645.
- Mervyn STONE (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, 36: 111–147. Barnard's comment (proposing the vote of thanks): 133–135.
- William F. STOUT (1970). A martingale analogue of kolmogorov's law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 15: 279–290.
- Alan STUART, Keith J. ORD, and Steven ARNOLD (1999). *Kendall's Advanced Theory of Statistics*, volume 2a: Classical Inference and the Linear Model. Arnold, London, sixth edition.
- William S. Gossett (STUDENT) (1908). On the probable error of a mean. *Biometrika*, 6:1–25.
- Richard S. SUTTON and Andrew G. BARTO (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Kei TAKEUCHI (1975). *Statistical Prediction Theory*. Baifukan, Tokyo. In Japanese.
- John W. TUKEY (1947). Nonparametric estimation II: Statistically equivalent blocks and tolerance regions – the continuous case. *Annals of Mathematical Statistics*, 18:529–539.
- Alan M. TURING (1950). Computing machinery and intelligence. *Mind*, 59:433–460.
- William T. TUTTE (2001). *Graph Theory*. Cambridge University Press, Cambridge, England.
- Leslie G. VALIANT (1984). A theory of the learnable. *Communications of the ACM*, 27:1134–1142.
- Vladimir N. VAPNIK (1982). *Estimation of Dependences Based on Empirical Data*. Springer, New York. This is the English translation of: Вапник Владимир Наумович, Восстановление зависимостей по эмпирическим данным, Nauka, Moscow, 1979.
- Vladimir N. VAPNIK (1995). *The Nature of Statistical Learning Theory*. Springer, New York. Second edition: 2000.
- Vladimir N. VAPNIK (1998). *Statistical Learning Theory*. Wiley, New York.
- Vladimir N. VAPNIK and Aleksei Ya. CHERVONENKIS (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Soviet Mathematics Doklady*, 9:915–918.

- Vladimir N. VAPNIK and Aleksei Ya. CHERVONENKIS (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16:264–280.
- Vladimir N. VAPNIK and Aleksei Ya. CHERVONENKIS (1974). Вапник Владимир Наумович и Червоненкис Алексей Яковлевич. Теория распознавания образов (*Theory of Pattern Recognition*). Nauka, Moscow. German translation: W. Vapnik and A. Tscherwonenskis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin.
- Vladimir N. VAPNIK and Alexander M. STERIN (1977). Ordered minimization of total risk in a pattern-recognition problem. *Automation and Remote Control*, 10: 1495–1503. Russian original in: *Автоматика и телемеханика*, 10:83–92.
- John VENN (1866). *The Logic of Chance*. Macmillan, London.
- Jean VILLE (1939). *Étude critique de la notion de collectif*. Gauthier-Villars, Paris.
- Vladimir VOVK (1986). On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248.
- Vladimir VOVK (1990). Aggregating strategies. In Mark Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA. Morgan Kaufmann.
- Vladimir VOVK (1993). A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society, Series B*, 55:317–351.
- Vladimir VOVK (1999). Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282.
- Vladimir VOVK (2001a). Competitive on-line statistics. *International Statistical Review*, 69:213–248.
- Vladimir VOVK (2001b). Probability theory for the Brier game. *Theoretical Computer Science*, 261:57–79.
- Vladimir VOVK (2002a). Asymptotic optimality of Transductive Confidence Machine. In *Proceedings of the Thirteenth International Conference on Algorithmic Learning Theory*, volume 2533 of *Lecture Notes in Artificial Intelligence*, pages 336–350, Berlin. Springer.
- Vladimir VOVK (2002b). On-line Confidence Machines are well-calibrated. In *Proceedings of the Forty-Third Annual Symposium on Foundations of Computer Science*, pages 187–196, Los Alamitos, CA. IEEE Computer Society. This paper's main results were first reported at the NeuroCOLT workshop in May 2002.
- Vladimir VOVK (2003a). Universal well-calibrated algorithm for on-line classification. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines: Sixteenth Annual Conference on Learning Theory and Seventh Kernel Workshop, COLT/Kernel 2003*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 358–372, Berlin. Springer. Appears in revised form in the Special Issue on Advances in Learning Theory and Kernel Methods of *Journal of Machine Learning Research* devoted to COLT/Kernel'2003, 5:575–604.
- Vladimir VOVK (2003b). Well-calibrated predictions from on-line compression models. In Ricard Gavaldà, Klaus P. Jantke, and Eiji Takimoto, editors, *Proceedings of the Fourteenth International Conference on Algorithmic Learning Theory*, volume 2842 of *Lecture Notes in Artificial Intelligence*, pages 268–282, Berlin. Springer.

- Vladimir VOVK (2004). Well-calibrated predictions from on-line compression models. This is the journal version (containing a section on hypergraphical models) of Vovk 2003b; to appear in the Special Issue of *Theoretical Computer Science* devoted to the ALT'2003 conference.
- Vladimir VOVK, Alex GAMMERMAN, and Craig SAUNDERS (1999). Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, San Francisco, CA. Morgan Kaufmann.
- Vladimir VOVK, David LINDSAY, Ilia NOURETDINOV, and Alex GAMMERMAN (2003a). Mondrian confidence machine, On-line Compression Modelling project, <http://vovk.net/kp>, Working Paper #4.
- Vladimir VOVK, Ilia NOURETDINOV, and Alex GAMMERMAN (2003b). Testing exchangeability on-line. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning*, pages 768–775, Menlo Park, CA. AAAI Press.
- Vladimir VOVK and Glenn SHAFER (2003). Kolmogorov's contributions to the foundations of probability. *Problems of Information Transmission*, 39:21–31.
- Vladimir VOVK and Vladimir V. V'YUGIN (1993). On the empirical validity of the Bayesian method. *Journal of the Royal Statistical Society, Series B*, 55:253–266.
- Vladimir V. V'YUGIN (1994). Algorithmic entropy (complexity) of finite objects and its applications to defining randomness and amount of information. *Selecta Mathematica Sovietica*, 13:357–389.
- Grace WAHBA (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.
- Abraham WALD (1943). An extension of Wilks' method for setting tolerance limits. *Annals of Mathematical Statistics*, 14:45–55.
- Abraham WALD (1947). *Sequential Analysis*. Wiley, New York.
- Abraham WALD and Jacob WOLFOWITZ (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19:326–339.
- Chris WATKINS and Jason WESTON (1999). Support vector machines for multiclass pattern recognition. In *Proceedings of the Seventh European Symposium on Artificial Neural Networks*, pages 219–224.
- Samuel S. WILKS (1941). Determination of samples sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12:91–96.
- David WILLIAMS (1991). *Probability with Martingales*. Cambridge University Press, Cambridge, England.

Index

Symbols

- \mathcal{F} -conformal predictor 261
 σ -algebra 275
 Borel 276
 generated by a random element 276
 independent σ -algebras 277
1-inclusion graph 271

A

- adapted sequence of random elements 284
agreement
 between a probability distribution and a parametric detrending transformation 181
 between a probability distribution and an on-line compression model 191, 194
 between an example and a configuration 226
 between an example and a summary 226
 between configurations 229
algorithmic randomness 215, 218, 219
 discarded 50, 138, 216
algorithmic randomness deficiency 49, 218
 Kolmogorov *see* Kolmogorov deficiency
 repetitive *see* repetitive deficiency
algorithmic randomness level 218
almost certain event 276
annealed entropy 250

- Arbuthnott, John (1667–1735) 168, 254
assumption
 exchangeability 18
 randomness 2
asymmetric classification 114, 115
asymptotic efficiency 53
attribute 231

B

- backward kernel 190
bag 23
bag σ -algebra 283
Bayes error 68
Bayes, Thomas (1701–1761) 264
Bayesian learning 264
Bernoulli problem 159
Bernoulli sequence 215
Bernoulli, Jacob (1654–1705) 241, 243–245, 272
 Ars Conjectandi 243
 Bernoulli's theorem 243
boosting 104, 130
bootstrapping cases 130
bootstrapping residuals 130
bootstrap 130
bootstrap samples 103
Borel σ -algebra 276
Borel space 276
Boston Housing data set 39–42, 204, 267, 292

C

- calibration curve 68

- predictable 78
 calibration event 150
 calibration martingale 150
 calibration set 111
 calibration supermartingale 150
 category 114, 158, 159, 198, 224
 centered modified residuals 103
 central limit theorem 244
 classification 2, 9
 cluster 225
 complete statistical test 136
 conditional expectation 279
 version of 279
 conditional probability 280
 conditional validity 114
 confidence 96, 162
 confidence level 8, 19
 confidence predictor 8, 19
 (δ_n) -conservative 112
 asymptotically conservative 20, 21,
 60, 68
 asymptotically conservative in
 probability 108
 asymptotically exact 20, 21
 asymptotically exact in probability
 108
 asymptotically optimal 68
 at least as good as another predictor
 42
 category-wise conservative 115
 category-wise exact 115
 category-wise valid 115
 conservatively valid 21
 deterministic *same as* confidence
 predictor
 exactly valid 21
 invariant 42, 68, 108
 normalized 46
 randomized *see* randomized
 confidence predictor
 weakly exact 47
 configuration 225
 conformal prediction 7
 conformal predictor 26, 193, *see also*
 nonconformity measure
 \mathcal{L} -taught 107
 1-nearest neighbor 56
 determined by a nonconformity
 measure 26
 deterministic *same as* conformal
 predictor
 dynamic 181
 inductive 98
 inductive smoothed 99
 NNR 38
 off-line 110
 smoothed *see* smoothed conformal
 predictor
 conformal transducer 44, 192
 as Mondrian transducer 117
 smoothed *see* smoothed conformal
 transducer
 conformity 10
 conformity measure 24, *see also*
 nonconformity measure
 conformity score 24, *see also* noncon-
 formity measure
 conservative validity 9, 20
 for confidence predictors 21
 for transducers 45
 consistent table set 226
 convergence
 almost surely 277
 in probability 277
 count 226
 Cournot's principle 148
 credibility 96
 critical significance level 67, 70, 71, 74
 cumulative lower error probability curve
 161
 cumulative upper error probability
 curve 161
- D**
- data sequence 18
 compatible with a function class
 261
 incomplete 19
 Kolmogorov-typical 219
 repetitive-typical 219
 De Moivre, Abraham (1667–1754)
 243–245
 De Moivre's theorem 244
 decision tree 130
 deductive step 6
 deleted LSCM 34
 deleted prediction 28
 deleted residual 34

- detrended example space 181
 digraph 206
 discrepancy measure 28
 discriminant analysis 95
 distance 62
 distribution of a random element 276
 diverse data set 13, 132
 domination in distribution 21
 Doob, Joseph L. (1910–2004) 171
 dual form of ridge regression 35
 dynamic conformal predictor 181
 dynamic model based on exchangeability 181
- E**
 efficiency 9, 148
 empirical calibration curve 64
 empirical risk minimization 247
 emptiness curve 66
 empty prediction 53
 empty summary 190
 error probability interval 161
 Eulerian path 206
 event 276, *see also* game event
 almost certain 276
 exact validity 9
 exactly valid randomized transducer 44
 example 2, 17, 225, *see also* extended example
 example space 18
 exchangeability model 196, 282
 exchangeability supermartingale 170
 exchangeable probability distribution 18, 282
 extended example 61
 n-conforming 86
 n-strange 86
 extended object 22, 61
 extended object space 22
 extended random variable 276
- F**
 factorial-product 229
 FCVP *see* fully conditional Venn predictor
 feature space 36
 fiducial probability distribution 255
 filtration 284
- finitary repetitive structure 225
 Fisher, Ronald A. (1890–1962) 241, 254–256
 forward function 190
 frame 225
 fully conditional taxonomy 227
 fully conditional Venn predictor 227
- G**
 game event 148
 calibration event 150
 game martingale 150, 157
 calibration martingale 150, 158
 reversible martingale 156, 158
 game space 148
 game supermartingale 150, 157
 calibration supermartingale 150, 158
 reversible supermartingale 156, 158
 Gauss linear model 201
 Gauss linear shuffling 204
 Gaussian model 199
 genetic algorithms 130
 ghost predictor 123
 Gosset, William S. *see* Student
- H**
 hat matrix 31
 hedged prediction XV, 5
 high-dimensional environment 3
 highest probability density region 102
 horizon 145, 194, 224
 hypergraphical model 225
 reduced 228
 hypergraphical structure 225
- I**
 ICP *see* inductive conformal predictor
 ideal teacher 7, 107
 idempotent matrix 31
 incomplete data sequence 19
 incomplete gamma function 172
 incomplete statistical test 135
 increasing sequence of σ -algebras 284
 increasing sequence of random variables 284
 indicator function 59
 inductive conformal predictor 98
 1-nearest neighbor 101

- off-line 111
- semi-off-line 112
- smoothed 99
- inductive learning 5, 242
- inductive prediction 6
 - deductive step 6
 - inductive step 6
- inductive-exchangeability model 197
- inductivist objection 262
- inequality
 - Doob's 170, 285, 290
 - Hoeffding's 287, 290
 - Hoeffding–Azuma 290
- invariant confidence predictor 42, 68, 108
- invariant simple predictor 28
- J**
- junction tree 228
- K**
- kernel 36
- Kolmogorov complexity 214
- Kolmogorov complexity model 214
- Kolmogorov deficiency 219
- Kolmogorov, Andrei Nikolaevich (1903–1987) XVI, 15, 49, 140, 189, 214, 290
- L**
- label 2, 17, 225, 231
- label space 17, 231
- label variable 231
- Laplace's rule of succession 159, 263
- Laplace, Pierre Simon (1749–1827) 264
- law of the iterated logarithm 286
- lazy teacher 107
- learning
 - Bayesian 264
 - inductive 5, 242
 - off-line 5
 - on-line 5
 - transductive 5, 253
 - under randomness 3
- least squares confidence machine 34
 - deleted 34
 - studentized 34
- least squares regression 30
- lemma
 - Borel–Cantelli 285
 - Borel–Cantelli–Lévy 285
- level δ statistical test 149
- level δ martingale test 152
- Lévy, Paul (1886–1971) 290
- lexicographic order 62
- logistic regression 106
- LSCM *see* least squares confidence machine
- M**
- marginalization 229
- Markov graph 208
- Markov kernel 139, 280
- Markov sequence 215
- Markov summary 207
- Martin-Löf, Per (born 1942) 138, 215
- martingale 284
- martingale difference 284
- martingale predictor 252
 - valid 252
- martingale test 152
- MCP *see* Mondrian conformal predictor
- MCT *see* Mondrian conformal transducer
- measurable function 276
- measurable partition 158
- measurable set 275
- measurable space 275
 - Borel 276
- Mises, Richard von (1883–1953) 168, 217
- model
 - exchangeability 282
 - randomness 278
- Mondrian conformal predictor 116
- Mondrian conformal transducer 115
- Mondrian nonconformity measure 114
- Mondrian taxonomy 114, 198
 - equivalent 117
 - more general 116
- Mondrian, Piet (1872–1944) 116
- Mondrian-exchangeability model 198
- multilabel classification 58
 - one-against-one procedure 59
 - one-against-the-rest procedure 59
- multiple prediction 53

- multiplicity curve 66
multiprobability predictor 156
valid 158, 224
Venn predictor 159, 224
multiset 23
- N**
natural table 225
natural table set 226
nearest neighbors regression 38
nearest neighbors smoothed conformal predictor 63
neural networks 130
NN power martingale 173
NN SJ martingale 177
NN SJ supermartingale 177
NN SM martingale 173
NN SM supermartingale 173
NN transducer 173
NNR *see* nearest neighbors regression
NNR conformal predictor 38
nonconformity measure 23, 192
from boosting 105
from bootstrap 103
from de-Bayesing 102
from decision trees 104
from least squares 30
from logistic regression 106
from nearest neighbors
classification 54, 63, 101
regression 38
from neural networks 105
from ridge regression 30
from support vector machines 58, 95
general scheme 28, 54
nonconformity score 24, *see also* nonconformity measure
normalization 292
normalized confidence predictor 46
- O**
object 2, 17, 231, *see also* extended object
absent 18
object space 17, 231
OCM *see* on-line compression model
off-line conformal predictor 110
- off-line inductive conformal predictor 111
off-line learning 5
on-line compression model 15, 190, *see also* repetitive structure
exchangeability 196
Gauss linear 201
Gaussian 199
hypergraphical 226
inductive-exchangeability 197
Markov 205
Mondrian-exchangeability 198
reduced 192
with finite horizon 194
on-line learning 5
one-against-one procedure 59
one-against-the-rest procedure 59
optimal separating hyperplane 58
- P**
p-value 25, 44, 45, 192
smoothed 27
PAC theory 248
PAC transduction 260
parameter space 276
parametric detrending transformation 181
Pearson, Karl (1857–1936) 244
performance curve 68
predictable 78
power martingale 172
power probability distribution 18, 278
power probability space 278
power supermartingale 172
predictability 65
predictability distribution function 65
predictable calibration curve 78
predictable performance curve 78
predictable sequence of random elements 284
prediction 96, *see also* prediction set
Bayesian 264
hedged XV, 5
inductive 242
probabilistic 12
probably approximately correct 4
simple 3
transductive 253
prediction rule 3

- prediction set 8
 empty 53
 multiple 53
 singular 80
 prediction space 28
 prior distribution 242
 probabilistic prediction 12
 probabilistic predictor 139, 154
 N -calibrated 156
 strongly N -calibrated 156
 universally consistent 12
 weakly N -calibrated 155
 probability 276
 upper 149
 probability distribution 276
 compatible with a function class 261
 exchangeable 18, 282
 regular 69
 probability estimator 133, 134
 weakly valid 133, 134
 probability forecasting 145
 probability space 276
 problem of the reference class 159
 proper training set 111
- R**
- random element 276
 independent random elements 277
 random noise 35, 177
 random variable 276
 extended 276
 randomized confidence predictor 22
 asymptotically conservative for... 60
 asymptotically optimal 61
 exactly valid 23
 strongly exact 45
 universal 61
 randomized exchangeability martingale 171
 randomized transducer 44, 192
 exactly valid 44, 192
 randomness assumption 2, 18
 randomness model 278
 randomness supermartingale 171
 reduced hypergraphical model 228
 reduction of classification to regression 54
- regression 2, 9
 least squares 30
 logistic 106
 nearest neighbors 38
 ridge regression 29
 regular conditional distribution 282
 regular conditional probability 281
 regular probability distribution 69
 rejection 135
 repetitive algorithmic randomness
 deficiency 219
 repetitive structure 195, *see also*
 on-line compression model
 exchangeability 196
 finitary 225
 Gauss linear 201
 Gaussian 199
 hypergraphical 226
 inductive-exchangeability 197
 Markov 205
 Mondrian-exchangeability 198
 more general 197
 more specific 197, 225
 reduced 195
 reproducing kernel Hilbert space 51
 reshuffling a data set 293
 residual 30
 reversible martingale 156
 reversible supermartingale 156
 ridge parameter 30
 ridge regression 29
 dual form 35
 ridge regression confidence machine 32
 RRCM *see* ridge regression confidence machine
- S**
- sample space 276
 semi-Markov graph 206
 semi-off-line inductive conformal
 predictors 112
 separator 228
 sequence of random elements
 adapted 284
 predictable 284
 shuffling a data set 294
 significance level 19, 149
 critical *see* critical significance level

- simple mixture 172
 simple prediction 3
 simple predictor 18
 invariant 28
 singular prediction 80
 size of a table 225
 size of a table set 226
 Sleepy Jumper 176
 Sleepy Jumper martingale 177
 Sleepy Jumper supermartingale 177
 slow teacher 107, 122
 smoothed conformal predictor 27
 \mathcal{L} -taught 107
 smoothed conformal transducer 44, 192
 smoothed inductive conformal predictor 99
 smoothed Mondrian conformal transducer 115
 smoothed p-value 27
 state 208
 statistical learning theory 3
 statistical model 214, 242, 276
 statistical test 149, 218
 complete 136
 incomplete 135
 uniform *see* uniform test
 strong law of large numbers
 Borel's 286
 Kolmogorov's 286
 martingale 286
 structural risk minimization 249, 260
 Student (William S. Gosset, 1876–1937) 254–256
 Student predictor 202
 studentized LSCM 34
 sub- σ -algebra 275
 summary (in a hypergraphical model) 226
 summary space 190, 195
 supermartingale 169, 284
 P -supermartingale 170
 support vector 95
 support vector machine 56, 95
 dual problem 58
 SVM *see* support vector machine
 symmetric classification 120
- T**
 table 225
 natural 225
 size of a table 225
 table set 225
 consistent 226
 empty 226
 generated by a data sequence 226
 natural 226
 size of a table set 226
 target label 231
 target label space 231
 target label variable 231
 taxonomy *same as* Venn taxonomy;
 not to be confused with Mondrian taxonomy
 teacher
 ideal 7, 107
 lazy 107
 slow 107
 teaching schedule 107
 test set 4
 theorem
 de Finetti's 283
 Glibenko–Cantelli 248
 Littlestone and Warmuth's 249
 McDiarmid's 288, 290
 Ville's 164, 285, 290
 time series 203
 tolerance predictor 256, 257
 training set 3, 4, 258
 transducer 44, 192
 conservatively valid 45
 deterministic *same as* transducer
 randomized *see* randomized transducer
 transductive learning 5, 253
 transductive prediction 6
 trial 9, 18
 update 98
 Turing, Alan M. (1912–1954) 1, 15
 typicalness *see* algorithmic randomness
- U**
 unconstrained randomness 3
 underlying algorithm 11
 uniform law of large numbers 246
 uniform test 218

- universal 218
 - universal predictor 61
 - universal Turing machine 49
 - universal uniform test 218
 - universally consistent probabilistic predictor 12
 - update trial 98
 - upper probability 149
 - USPS data set 2, 3, 10, 55–57, 64, 70, 72, 73, 101, 117–119, 121, 161, 162, 172–175, 178, 249, 250, 291
 - exact 9, 20
 - variable 225
 - variation distance 163
 - VC dimension 246
 - Venn predictor 159, 224
 - Venn taxonomy 158, 224
 - fully conditional 227
 - Venn, John (1834–1923) 168
 - version of conditional expectation 279
 - Ville, Jean (1910–1988) 168, 290
 - Ville’s theorem 164, 285, 290
- V**
- validity 9, 158, 224
 - asymptotic 20
 - conditional 114
 - conservative 9, 20
 - weak learner 104
 - weak validity 133, 134
 - working example 110
 - working set 258
- W**