THEORY AND DECISION LIBRARY
SERIES B: MATHEMATICAL AND STATISTICAL METHODS

Eyke Hüllermeier

# Case-Based Approximate Reasoning

Springer

# CASE-BASED APPROXIMATE REASONING

THEORY AND DECISION LIBRARY

General Editors: W. Leinfellner (Vienna) and G. Eberlein (Munich)

Series A: Philosophy and Methodology of the Social Sciences

Series B: Mathematical and Statistical Methods

Series C: Game Theory, Mathematical Programming and Operations Research

SERIES B: MATHEMATICAL AND STATISTICAL METHODS

VOLUME 44

*Scope:* The series focuses on the application of methods and ideas of logic, mathematics and statistics to the social sciences. In particular, formal treatment of social phenomena, the analysis of decision making, information theory and problems of inference will be central themes of this part of the library. Besides theoretical results, empirical investigations and the testing of theoretical models of real world problems will be subjects of interest. In addition to emphasizing interdisciplinary communication, the series will seek to support the rapid dissemination of recent results.

*The titles published in this series are listed at the end of this volume.*

# CASE-BASED APPROXIMATE REASONING

*by*

EYKE HÜLLERMEIER
*University of Magdeburg, Germany*

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-5694-X (HB)
ISBN-13 978-1-4020-5694-9 (HB)
ISBN-10 1-4020-5695-8 (e-book)
ISBN-13 978-1-4020-5695-6 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

*www.springer.com*

*Printed on acid-free paper*

To Christiane, Jana, and Jaël

# Table of Contents

**Dedication** . . . . . . v

**Foreword** . . . . . . xiii

**Preface** . . . . . . xv

**Notation** . . . . . . 1

**1. Introduction** . . . . . . 5

1.1 Similarity and case-based reasoning . . . . . . 5

1.2 Objective of this book . . . . . . 6

1.2.1 Making case-based inference more reliable . . . . . . 9

1.2.2 The important role of models . . . . . . 10

1.2.3 Formal models of case-based inference . . . . . . 11

1.3 Overview . . . . . . 13

**2. Similarity and Case-Based Inference** . . . . . . 17

2.1 Model-based and instance-based approaches . . . . . . 17

2.1.1 Model-based approaches . . . . . . 18

2.1.2 Instance-based approaches . . . . . . 19

2.1.3 Knowledge representation . . . . . . 19

2.1.4 Performance in generalization . . . . . . 20

2.1.5 Computational complexity . . . . . . 21

2.2 Similarity-based methods . . . . . . 22

2.2.1 Nearest neighbor (NN) estimation . . . . . . 22

2.2.2 Instance-based learning . . . . . . 29

2.2.3 Case-based reasoning . . . . . . 30

2.3 The concept of similarity ..... 33
2.3.1 Similarity in case-based reasoning ..... 33
2.3.2 Similarity and fuzzy sets ..... 37
2.3.3 Aggregation of local similarity measures ..... 38
2.4 Case-based inference ..... 41
2.4.1 Deterministic inference problems ..... 45
2.4.2 Non-deterministic inference problems ..... 48
2.4.3 Formal models of case-based inference ..... 52
2.5 Summary and remarks ..... 53

**3. Constraint-Based Modeling of Case-Based Inference** ..... 59
3.1 Basic concepts ..... 59
3.1.1 Similarity profiles and hypotheses ..... 59
3.1.2 Generalized similarity profiles ..... 62
3.2 Constraint-based inference ..... 65
3.2.1 A constraint-based inference scheme ..... 65
3.2.2 Non-deterministic problems ..... 69
3.3 Case-based approximation ..... 72
3.3.1 Properties of case-based approximation ..... 73
3.3.2 Local similarity profiles ..... 79
3.4 Learning similarity hypotheses ..... 81
3.4.1 The learning task ..... 81
3.4.2 A learning algorithm ..... 85
3.4.3 Properties of case-based learning ..... 86
3.4.4 Experimental results ..... 94
3.5 Application to statistical inference ..... 97
3.5.1 Case-based parameter estimation ..... 97
3.5.2 Case-based prior elicitation ..... 98
3.6 Summary and remarks ..... 99

**4. Probabilistic Modeling of Case-Based Inference** ..... 103
4.1 Basic probabilistic concepts ..... 106
4.1.1 Probabilistic similarity profiles and hypotheses ..... 107

4.1.2 Generalized probabilistic profiles . . . 110

4.2 Case-based inference, probabilistic reasoning, and statistical inference 112

4.3 Learning probabilistic similarity hypotheses . . . 117

4.3.1 Simple hypotheses and credible case-based inference . . . 117

4.3.2 Extended case-based learning . . . 118

4.4 Experiments with regression and label ranking . . . 121

4.4.1 Regression: artificial data . . . 122

4.4.2 Regression: real-world data . . . 124

4.4.3 Label ranking . . . 126

4.5 Case-based inference as evidential reasoning . . . 129

4.5.1 Transformation of probabilistic evidence . . . 131

4.5.2 Inference from individual cases . . . 133

4.5.3 Combining evidence from several cases . . . 134

4.6 Assessment of cases . . . 140

4.6.1 Similarity-weighted approximation . . . 140

4.6.2 More general criteria . . . 142

4.6.3 Assessment of individual cases . . . 144

4.7 Complex similarity hypotheses . . . 146

4.7.1 Inference schemes of higher order . . . 147

4.7.2 Partially admissible profiles . . . 150

4.8 Approximate probabilistic inference . . . 152

4.8.1 Generalized uncertainty measures and profiles . . . 152

4.8.2 An approximate inference scheme . . . 155

4.9 Summary and remarks . . . 159

**5. Fuzzy Set-Based Modeling of Case-Based Inference I . . . 165**

5.1 Background on possibility theory . . . 165

5.1.1 Possibility distributions as generalized constraints . . . 166

5.1.2 Possibility as evidential support . . . 168

5.2 Fuzzy rule-based modeling of the CBI hypothesis . . . 169

5.2.1 Possibility rules . . . 170

5.2.2 Modeling the CBI hypothesis . . . 171

5.3 Generalized possibilistic prediction . . . 176

5.3.1 Control of compensation and accumulation of support ...... 177
5.3.2 Possibilistic support and weighted NN estimation .......... 178
5.3.3 Upper and lower possibility bounds ...................... 179
5.3.4 Fuzzy logical evaluation ................................ 180
5.3.5 Comparison of extrapolation principles .................. 181
5.3.6 From predictions to decisions ........................... 186
5.3.7 An illustrative example ................................. 188
5.3.8 Complexity issues ....................................... 191
5.4 Extensions of the basic model ............................... 191
5.4.1 Dealing with incomplete information ..................... 192
5.4.2 Discounting noisy and atypical instances ................ 194
5.4.3 From instances to rules ................................. 198
5.4.4 Modified possibility rules ............................... 200
5.4.5 Combination of several rules ............................. 202
5.4.6 Locally restricted extrapolation ......................... 204
5.4.7 Incorporation of background knowledge ................... 205
5.5 Experimental studies ........................................ 206
5.5.1 Preliminaries ............................................ 206
5.5.2 Classification accuracy .................................. 207
5.5.3 Statistical assumptions and robustness .................. 209
5.5.4 Variation of the aggregation operator ................... 210
5.5.5 Representation of uncertainty ........................... 211
5.6 Calibration of CBI models ................................... 212
5.7 Relations to other fields .................................... 219
5.7.1 Fuzzy and possibilistic data analysis .................... 219
5.7.2 Fuzzy set-based approximate reasoning .................. 222
5.8 Summary and remarks ........................................ 224

**6. Fuzzy Set-Based Modeling of Case-Based Inference II** ......... 229
6.1 Gradual inference rules ..................................... 230
6.1.1 The basic model ......................................... 230
6.1.2 Modification of gradual rules ............................ 232
6.2 Certainty rules .............................................. 235

6.3 Cases as information sources . . . . . . 239

6.3.1 A probabilistic model . . . . . . 239

6.3.2 Combination of information sources . . . . . . 241

6.4 Exceptionality and assessment of cases . . . . . . 243

6.5 Local rules . . . . . . 247

6.6 Summary and remarks . . . . . . 249

**7. Case-Based Decision Making** . . . . . . 253

7.1 Case-based decision theory . . . . . . 255

7.2 Nearest Neighbor decisions . . . . . . 258

7.2.1 Nearest Neighbor classification and decision making . . . . . . 258

7.2.2 Nearest Neighbor decision rules . . . . . . 260

7.2.3 An axiomatic characterization . . . . . . 261

7.3 Fuzzy modeling of case-based decisions . . . . . . 265

7.3.1 Basic measures for act evaluation . . . . . . 266

7.3.2 Modification of the basic measures . . . . . . 267

7.3.3 Interpretation of the decision criteria . . . . . . 268

7.4 Fuzzy quantification in act evaluation . . . . . . 270

7.5 A CBI framework of CBDM . . . . . . 274

7.5.1 Generalized decision-theoretic setups . . . . . . 274

7.5.2 Decision making using belief functions . . . . . . 277

7.5.3 Possibilistic decision making . . . . . . 279

7.6 CBDM models: A discussion of selected issues . . . . . . 282

7.6.1 The relation between similarity, preference, and belief . . . . . . 282

7.6.2 The effect of observed cases . . . . . . 285

7.6.3 Dynamic aspects of decision making . . . . . . 286

7.7 Experience-based decision making . . . . . . 287

7.7.1 Compiled decision models . . . . . . 290

7.7.2 Satisficing decision trees . . . . . . 292

7.7.3 Experimental evaluation . . . . . . 296

7.8 Summary and remarks . . . . . . 301

**8. Conclusions and Outlook** . . . 307

**A. Possibilistic Dominance in Qualitative Decisions** . . . 309

**B. Implication-Based Fuzzy Rules as Randomized Gradual Rules** 313

B.1 Implication-based fuzzy rules . . . 314

B.1.1 Gradual rules . . . 315

B.1.2 Other implication-based rules . . . 317

B.2 Randomized gradual rules . . . 320

B.3 A probabilistic representation of implication-based fuzzy rules . . . 322

**C. Similarity-Based Reasoning as Logical Inference** . . . 333

**D. Simulation Results of Section 3.4.4** . . . 335

**E. Experimental Results of Section 5.5.4** . . . 339

**F. Simulation Results of Section 7.4** . . . 341

**G. Computation of an Extended Splitting Measures** . . . 343

**H. Experimental Results of Section 7.7.2** . . . 345

**References** . . . 346

**Index** . . . 369

# Foreword

In the last three decades, important trends of research in artificial intelligence have been devoted to the design and the study of inference systems that exhibit human-like reasoning capabilities, with special emphasis on tolerance toward incomplete information and uncertainty. This program, which can be also related to cognitive psychology concerns, has led to the development of computational models, in particular for default reasoning that accommodates exceptions and inconsistency, for approximate reasoning with interpolative abilities, as well as for case-based reasoning.

Case-based reasoning (CBR) relies on the concept of similarity, and more particularly on the idea that situations recognized as similar in important aspects may be also similar in other respects. CBR thus appears as a yet simple instance of analogical reasoning, but as powerful as a general problem solving method. This explains the success encountered by case-based reasoning, and beyond that, the interest for similarity-based reasoning that has been increasing in the last ten years. The very idea of CBR is thus to solve new problems on the basis of experience that is represented by already solved problems of the same type, referred to as cases. Thus, a new problem is solved by adapting the solution of a similar case, hoping that the adaptation can be done with much less effort than solving the problem from scratch.

A CBR system requires efficient techniques for several important subtasks, such as organizing and maintaining the case base, retrieving cases (which are maximally similar to the problem) from the case base, and adapting stored cases to the problem at hand. The basic inference mechanism underlying CBR, where the concept of similarity plays a major role, is built upon the principle of instance-based learning and nearest neighbor classification.

CBR has always been motivated by real-world problems, and research in this area has largely focused on building efficient computer systems. Less work has been done, however, on the theoretical foundations of case-based and similarity-based inference. Note that this strongly contrasts with the situation in default reasoning, and to a less extent in approximate reasoning. The present book seeks to remedy this flaw. Its objective is to contribute to a sound foundation of CBR and related fields, such as instance-based learning and analogical reasoning, by providing formal models of similarity-based inference.

To accomplish this objective, Eyke Hüllermeier embeds case-based inference into different frameworks of knowledge representation and reasoning, namely constraint-based reasoning, probability theory, and fuzzy sets and possibility theory. His basic idea is to express the heuristic “similar problem-similar solution” principle in the form of an explicit model, using the formal language of the underlying framework. Proceeding this way, one can take advantage of the reasoning mechanisms offered by that framework. Thus, the author develops various alternative methods, realizing case-based inference as constraint propagation, as probabilistic inference, and as fuzzy set-based approximate reasoning. As convincingly shown in the book, this approach has several advantages. Particularly, case-based inference can benefit from the features offered by the different frameworks. As a noticeable example consider the issue of uncertainty representation. How reliable is a solution proposed by a CBR system? Questions of such kind, which have not received much attention in CBR so far, can adequately be approached by means of probabilistic methods. Likewise, a fuzzy set-based approach to case-based inference can adequately cope with imprecisely or vaguely described cases. Needless to say, the alternative formalizations of similarity-based inference developed by the author are complementary rather than competitive, and different applications will usually call for different methods.

Apart from contributing to the formal foundations of case-based inference, the author’s approach has further advantages. Integrating case-based reasoning with other computing paradigms such as probabilistic reasoning and fuzzy set theory can lead to efficient hybrid methods and flexible information processing systems. This can help to clarify differences between alternative methods, but also to show things they have in common. In fact, as one of the more recent trends in artificial intelligence and machine learning, the development of hybrid systems has already produced a number of successful applications. A very promising approach in this connection is the combination of case-based and rule-based reasoning: Since cases and rules can represent, respectively, individual facts and generalized knowledge, these approaches can complement each other in a reasonable way. Indeed, close connections at a formal level can be observed between case-based inference and rule-based inference techniques as realized in approximate reasoning. Worth mentioning is also the current interest, reflected by the organization of specialized workshops, in the combination of case-based reasoning and techniques such as fuzzy sets, neural networks and genetic algorithms, often called “soft computing” paradigms.

With these motivations in mind, Eyke Hüllermeier has done an excellent job in writing this book. His monograph, which is the first one of this type, presents state of the art information as well as novel ideas and new research results in a highly readable and intelligible form. From a theoretical point of view, it clearly constitutes an important contribution to the foundations of case-based and approximate reasoning. From a practical point of view, it should be of interest to everyone working in CBR, fuzzy sets, uncertain reasoning, and related fields.

**Toulouse, July 2006** **Henri Prade**

# Preface

Longstanding research in artificial intelligence and related fields has produced a number of paradigms for building intelligent and knowledge-based systems, such as rule-based reasoning, constraint processing, or probabilistic graphical models. Being one of these paradigms, *case-based reasoning* (CBR) has received a great deal of attention in recent years and has been used successfully in diverse application areas.

The CBR methodology is inspired by human problem solving and has roots in cognitive psychology. Its key idea is to tackle new problems by referring to similar problems that have already been solved in the past. More precisely, CBR proceeds from individual experiences in the form of *cases*. The generalization beyond these experiences is largely founded on principles of analogical reasoning in which the (cognitive) concept of *similarity* plays an essential role.

This book is an attempt to contribute to the theoretical foundations of CBR, which are not as fully developed as one might expect in light of the practical success of the methodology. To this end, we propose formal models of the fundamental though often implicitly used inference principle underlying CBR methods, namely the heuristic assumption that “similar problems have similar solutions”. Proceeding from these models, concrete inference schemes are derived within different frameworks of approximate reasoning and reasoning under uncertainty.

The *case-based approximate reasoning* methods thus obtained especially emphasize the heuristic nature of case-based (similarity-based) inference. More specifically, we combine case-based reasoning with probabilistic methods as well as fuzzy set-based modeling and approximate reasoning techniques. This way, we hope to contribute to a solid foundation of case-based reasoning which is grounded on well-established concepts and techniques from the aforementioned fields, but also to inspire new approaches and to cast light on already existing ones.

Needless to say, the application of these reasoning methods is not restricted to CBR in a narrow sense. Instead, these methods suggest “case-based” approaches in other fields as well. In the final part of the book, we discuss models of *case-based decision making* which combine principles of both case-based reasoning and decision theory. Such models are motivated for reasons of cognitive plausibility as well as practical relevance, and can complement existing models, such as expected utility theory, in a reasonable way.

Much of the research underlying this book has been conducted during my stays at the *Institut de Recherche en Informatique de Toulouse* (IRIT) in the research group headed by Didier Dubois and Henri Prade. I would like to express my gratitude to both of them for providing this opportunity. The IRIT always offered an extremely stimulating research environment, and many of the ideas presented in this book emerged in discussions with Didier and Henri.

I am likewise indebted to my family who not only sustained the times of my absence and adopted the long hours that I have spent writing the manuscript, but also rendered every assistance and moral encouragement whenever needed. To them, Christiane, Jana, and Jaël, I dedicate this book.

Magdeburg, Germany Eyke Hüllermeier
March 2006

# Notation

This list offers some of the basic notation that will be used throughout the book. Specific notation will be introduced in the main text as the need arises. Even though some symbols will be used for different purposes, the concrete meaning should always be clear from the context.

## Basic mathematical notation

| | |
|---|---|
| $\forall$, $\exists$ | universal and existential quantifier |
| $\wedge$, $\vee$, $\neg$ | logical conjunction, disjunction, negation |
| $\Rightarrow$, $\Leftrightarrow$ | logical implication, equivalence |
| $\overset{\text{df}}{=}$, $\equiv$ | equality by definition, identity |
| $\mathfrak{N}$ ($\mathfrak{N}_0$) | set of natural numbers (including 0) |
| $\mathfrak{I}$ | set of integers |
| $\mathfrak{Q}$ | set (field) of rational numbers |
| $\mathfrak{R}$, $\overline{\mathfrak{R}}$ | set (field) of real numbers (including $-\infty$ and $\infty$) |
| $\mathfrak{R}_{\geq\alpha}$ $\mathfrak{R}_{>\alpha}$ | set of real numbers (equal to or) larger than $\alpha$ |
| $\aleph_0$ | cardinality of the natural numbers |
| card$(X)$, $\lvert X\rvert$ | cardinality of the set $X$ |
| $X \oplus Y$ | addition of sets: $X \oplus Y = \{x + y \mid x \in X, y \in Y\}$ |
| $\emptyset$ | empty set |
| $\cup, \cap, \setminus$ | set-theoretic union, intersection, difference |
| $X \times Y$ | Cartesian product of sets $X$ and $Y$ |
| $\subset$ resp. $\subseteq$, $\subsetneq$ | subset relation, proper subset relation |
| $\lvert\cdot\rvert$, $\lvert\cdot\rvert_p$ | Euclidean distance (in $\mathfrak{R}^n$), $\mathcal{L}_p$-norm |
| $\lVert f\rVert_p$ | $\mathcal{L}_p$-norm of the function $f$ |
| $[\alpha, \beta]$, $(\alpha, \beta]$, ... | closed and (half) open intervals |
| diam$(X)$ | diameter $\sup_{x,x'\in X} \mathit{\Delta}(x, x')$ of the set $X$ |
| $\mathfrak{B}_\varepsilon(x)$, $\bar{\mathfrak{B}}_\varepsilon(x)$ | open and closed $\varepsilon$-ball around $x$ |
| $(x_n)_{n\geq 1} \rightarrow x_0$ | convergence of the sequence $(x_n)_{n\geq 1}$ toward $x_0$ |
| $(x_n)_{n\geq 1} \nearrow x_0$ | convergence from below |
| $(x_n)_{n\geq 1} \searrow x_0$ | convergence from above |
| $O(f), o(f)$ | Landau symbols (with limit either 0 or $\infty$) |
| $\ll$ | much smaller than |
| dom$(f)$, rg$(f)$ | domain and range of the function $f$ |
| codom$(f)$ | codomain of the function $f$ |

| | |
|---|---|
| $f \wedge g$, $f \vee g$ | lower and upper envelope of the functions $f$ and $g$ |
| $f\|A$ | restriction of $f : X \longrightarrow Y$ to the set $A \subset X$ |
| $f \circ g$ | composition of functions $f$ and $g$ |
| $f \propto g$ | function $f$ is proportional to function $g$ |
| $\mathbb{I}_X$ | indicator function of the set $X$ |
| id | identical function $x \mapsto x$ |
| mod | modulo function (infix notation) |
| $\lfloor x \rfloor$ | largest integer $\leq x$ |
| max, min | maximum and minimum operator |
| inf, sup | infimum and supremum operator |
| exp | exponential function |
| ln, $\log_\alpha$ | natural logarithm, logarithm with a base $\alpha$ |
| arg max | $y \in \arg\max_{x \in X} f(x) \overset{\text{df}}{\Leftrightarrow} f(y) = \max_{x \in X} f(x)$ |
| $A = (a_{\imath\jmath}), a^{\jmath}$ | matrix $A$, $\jmath$-th column of $A$ |
| $A^t$ | transpose of matrix $A$ |
| $A \times x$, $x \times y$ | matrix and vector multiplication |
| $e_k$ | $k$-th unit vector in $\mathfrak{R}^n$ |

**Remark**: We use the term "increasing" (decreasing) in the strict sense, i.e., in the sense of strictly increasing (strictly decreasing). The opposite of increasing (decreasing), i.e., decreasing (increasing) in the weak sense, is referred to as "non-increasing" (non-decreasing). Even though the arg max-operator actually returns a *set* of elements, we shall often write $y = \arg\max_{x \in X} f(x) \in X$; it is then assumed that the maximizing element is unique, or that $y$ is simply an arbitrary choice among these elements.

## Probability theory and statistics

| | |
|---|---|
| $\mathcal{P}(\Omega, \mathcal{A})$, $\mathcal{P}(\Omega)$ | class of probability measures over the measurable space $(\Omega, \mathcal{A})$, $\mathcal{P}(\Omega) \overset{\text{df}}{=} \mathcal{P}(\Omega, 2^\Omega)$ |
| $\mathcal{F}(\Omega, \mathcal{A})$, $\mathcal{F}(\Omega)$ | class of normalized uncertainty measures over $(\Omega, \mathcal{A})$ resp. $(\Omega, 2^\Omega)$ |
| $\mu$, $\mu_{Y\|(X=x)}$ | probability measure, conditional measure |
| $\mathsf{Bel}$, $\mathsf{Pl}$, $\mathsf{m}$ | belief and plausibility function, mass distribution |
| $\mathbb{P}(X)$ | probability of event $X$ (informal notation) |
| $\otimes$ | product of measures |
| $\lambda(\cdot)$ | likelihood function |
| $\binom{n}{m}$ | binomial coefficient |
| $\preceq$ | stochastic dominance relation |
| $X \sim \mu$ | $X$ is distributed according to measure $\mu$ |

| | |
|---|---|
| $\phi_{\mu,\sigma}$ | density function of the normal distribution with mean $\mu$ and standard deviation $\sigma$ ($\phi \overset{\text{df}}{=} \phi_{0,1}$). |
| $\mu_{\varOmega}^{uni}$ | uniform measure over $\varOmega$ |
| $\mathsf{med}(A)$ | median of a set of numbers $A$ |
| $\mathbb{E}(X)$ | expected value of a random variable $X$ |
| $\mathbb{V}(X)$ | variance of a random variable $X$ |
| $\mathsf{bias}(\theta^e)$ | bias of an estimator $\theta^e$ |
| $\mathsf{MSE}(\theta^e)$ | mean square error of an estimator $\theta^e$ |

**Remark**: We shall often not distinguish between an element $x$ and its singleton $\{x\}$. In particular, we use the same notation for a probability measure and the related probability distribution function, i.e., we write $\mu(x)$ instead of $\mu(\{x\})$.

## Fuzzy sets, fuzzy measures, and possibility theory

| | |
|---|---|
| $\mathfrak{F}(X)$ | class of fuzzy subsets of a set $X$ |
| $A, B, \ldots$ | fuzzy sets (membership functions) |
| $\mu$ | membership function |
| $\mathrm{supp}(A)$ | support of the fuzzy set $A$ |
| $A_\alpha$, $A_0$ | $\alpha$-cut of $A$ ($0 < \alpha \leq 1$), closure of the support |
| $\top$, $\otimes$ | triangular norm (t-norm) |
| $\oplus$ | triangular co-norm (t-conorm) |
| $\rightsquigarrow$ | generalized (multiple-valued) implication operator |
| $\delta$, $\varDelta$ | possibility distribution (measure) |
| $\pi$, $\varPi$ | possibility distribution (measure) |
| $\mathcal{N}$ | necessity measure |
| $\int^{ch}$ | Choquet integral |
| $\int^{su}$ | Sugeno integral |

**Remark**: We do not distinguish between a fuzzy set and its membership function, i.e., we usually write $A(x)$ (rather than $\mu_A(x)$) for the degree of membership of an element $x$ in the fuzzy set $A$.

## Case-based inference

| | |
|---|---|
| $\mathcal{S}$, $\mathcal{R}$, $\mathcal{C} = \mathcal{S} \times \mathcal{R}$ | set of inputs, outputs, cases |
| $\mathcal{M}$, $\mathcal{M}^{\downarrow}$ | memory of cases $\langle s, r \rangle$, projection of $\mathcal{M}$ to $\mathcal{S}$ |
| $\varphi$ | mapping $\mathcal{S} \to \mathcal{R}$ or relation between $\mathcal{S}$ and $\mathcal{R}$ |
| $\widehat{\varphi}_{h,\mathcal{M}}$ | approximation of $\varphi$ based on similarity hypothesis $h$ and memory of cases $\mathcal{M}$ |
| $\sigma$, $\varDelta$ | similarity measure, distance measure |
| $D_{\mathcal{S}}$, $D_{\mathcal{R}}$ | ranges of the similarity measures $\sigma_{\mathcal{S}}$ (over $\mathcal{S}$) and $\sigma_{\mathcal{R}}$ (over $\mathcal{R}$) |
| $\varSigma$, $\langle \varSigma, s_0 \rangle$ | CBI setup, CBI problem with new input $s_0$ |
| $h_{\varSigma}$, $H_{\varSigma}$ | similarity profile, probabilistic similarity profile |
| $h$, $H$ | similarity hypothesis, probabilistic hypothesis |
| $\mathcal{N}_{\alpha}(r)$ | $\alpha$-neighborhood of an outcome $r \in \mathcal{R}$ |
| $\mathcal{N}_k(\mathcal{M}, s_0)$ | $k$-selection from memory $\mathcal{M}$ |
| $\mathcal{N}_k^{ex}(\mathcal{M}, s_0)$ | extended $k$-selection |
| $\mathsf{SST}(\mathcal{M}, s_0)$ | similarity structure |
| $\mathsf{pSST}(\mathcal{M}, s_0)$ | partial similarity structure |
| $\mathsf{OST}(\mathcal{M}, s_0)$ | outcome structure |
| $\mathsf{CST}(\mathcal{M}, s_0)$ | case structure |

## Abbreviations

| | |
|---|---|
| AI | artificial intelligence |
| CBA | case-based approximation |
| CBDM | case-based decision making |
| CBDT | case-based decision theory |
| CBI | case-based inference |
| CBL | case-based learning |
| CBLA, CBLP | case-based learning algorithm/process |
| CBR | case-based reasoning |
| EBDM | experience-based decision making |
| EUT | expected utility theory |
| ILP | integer linear programming |
| LS | least squares |
| ML, MLE | maximum likelihood (estimation) |
| PSP | probabilistic similarity profile |
| RCOP | repetitive combinatorial optimization |
| RSP | repeated search problem |

# 1. Introduction

## 1.1 Similarity and case-based reasoning

The idea that reasoning and problem solving (by human beings) are guided by experiences from situations which are similar to the current one has a long tradition in philosophy. It dates back at least to D. HUME who concedes to the concept of *similarity* (resemblance) the role of a basic principle of human thought:[1] "To me, there appear to be only three principles of connexion among ideas, namely, *Resemblance*, *Contiguity in time and place*, and *Cause and Effect*." Today, it is widely recognized that similarity plays a major role not only in commonsense but also in scientific reasoning. It is ubiquitous, for instance, in cognitive psychology, where it contributes essentially to formal theories of knowledge and behavior. It serves in particular as an organizing principle in object classification, generalization (e.g., extrapolation, interpolation), and the formation of concepts (e.g., clustering).

Interestingly enough, current research in ARTIFICIAL INTELLIGENCE (AI) seems to be largely dominated by the three principles stated by HUME.[2] Particularly, the idea that people reason by similarity (or, more generally, by analogy) has recently been realized on a formal level. In fact, the cognitive concept of similarity and the related everyday finding that – at least more often than not – similar causes bring about similar effects do now provide the basis of many formal methods of machine learning and (inductive) reasoning. As an example let us mention the NEAREST NEIGHBOR principle, a pattern recognition technique which proceeds from the assumption that "similar patterns belong to similar classes." In its basic version, this classifier estimates the class of a new instance (pattern) by the class of the "closest" among the already classified examples. Though it is worth mentioning that this rather simple decision principle provides the basis for more sophisticated methodologies, such as case-based reasoning (CBR).

CBR is a problem solving methodology built upon the hypothesis that "similar problems have similar solutions" [234]. Putting it in a somewhat more general

[1] See, e.g., [206], page 101 (our italics).

[2] The idea of a *cause–effect* relation, for instance, finds its formal counterpart in rule-based systems which have been (and still are) the most commonly used tool for realizing knowledge representation and reasoning (deductive inference) in knowledge-based systems. Moreover, the concept of *causality* is one of the central issues of current research in reasoning under uncertainty [337, 294]. Models of *temporal and spatial reasoning* have recently received much attention in the subfield of *qualitative reasoning* [242].

context – not necessarily related to special performance tasks such as classification or problem solving – we shall express assumptions of this kind, referring to a directed dependency between two entities, by saying that "similar inputs have similar outputs/outcomes." This hypothesis will be referred to as the CBI *hypothesis*, where CBI stands for *case-based inference*. Again, it is interesting to mention that the type of commonsense rule expressed by the CBI hypothesis goes back (at least) to HUME who notes that[3] "In reality, all arguments from experience are founded on the similarity, which we discover among natural objects, and by which we are induced to expect effects similar to those, which we have found to follow from such objects. ... From causes, which appear *similar*, we expect similar effects. This is the sum of all our experimental conclusions."

As can be seen, the concept of *similarity* is intimately related to the concept of a *case*, understood in a wide sense as a particular experience or "chunk" of knowledge: The very idea of case-based reasoning is to extrapolate such experiences to new situations that appear to be similar. Throughout this book, we shall often use the term "case-based reasoning" in this relatively broad sense, including the idea of case-based problem solving in the spirit of [234], that is, CBR in a more narrow sense, but also other case-based approaches, such as instance-based learning [11]. Thus defined, case-based reasoning is of course hardly conceivable without the idea of similarity. The other way round, however, the similarity concept is present in many other reasoning and learning methods as well, even though in a more implicit way. For example, the concept of a *kernel* in kernel-based learning methods [335, 339] can often be interpreted as a kind of similarity measure. For this reason, we will mostly speak about "case-based" rather than "similarity-based" reasoning (inference), even though in the context of this book the meaning of these two terms is more or less the same.

## 1.2 Objective of this book

Similarity disposes of a much less developed scientific tradition than other basic concepts of knowledge representation and reasoning, notably *preference* and *uncertainty*. The same remark does also apply to related reasoning techniques, despite the fact that, as noted above, many formal methods imply a sort of "representativeness assumption" which refers – at least implicitly – to similarity.[4] Yet how can the somewhat questionable reputation of similarity and case-based reasoning be explained? At least two reasons deserve mentioning.

The first point is related to the *concept* of similarity itself and concerns problems of *quantification* and *measurement*. In fact, similarity is extremely subjective and context dependent. Two things which are very similar from one point of view

[3] See e.g. [206], page 116.

[4] Function approximation by means of interpolation is a simple example.

can be very dissimilar from another perspective.[5] Moreover, it seems that similarity can be discovered between any pair of objects. Tversky [373] exemplifies this quite well by suspecting that "an essay is like a fish" (because both have head, body, and tail). In any case, a meaningful definition of similarity requires a considerable amount of *background knowledge*, e.g., for separating the relevant from the irrelevant features of an object, and strongly depends on the context and on the viewpoint of the observer. This might explain to some extent why similarity is considered with scepticism in modern scientific research that puts a strong emphasis on objectiveness and measurability.

Interestingly enough, the situation seems to be less delicate for the concepts of uncertainty and preference. As opposed to similarity, these concepts dispose of widely accepted formalizations. Uncertainty, for instance, can be formalized in terms of probability, which in turn can be connected with the concept of *frequency*. Undoubtedly, the latter can be grasped more easily than similarity. Probability can also be interpreted in a subjective sense, namely in terms of *fair betting rates*. Even if not being objective, this interpretation allows for the elicitation of probability degrees by means of a well-defined measurement procedure (at least in theory). The quantification of preference is a major concern of utility theory, by now a well-developed subfield of economic theory, and has more recently also received attention in AI [181, 96]. Again, utility is an extremely subjective concept. In many cases, however, it can be reduced to the concept of *cost* which is much more objective. Compared with similarity, it also disposes of better measurement devices.[6]

The second point is related to the *methods* of similarity-based reasoning. More precisely, it concerns the guiding principle underlying such methods, namely the above-mentioned CBI hypothesis. First, this hypothesis is apparently of heuristic nature. It should not be considered as a deterministic rule which is universally valid, but rather as a "rule of thumb" that can fail in some situations. In fact, similarity-based reasoning will lead to erroneous conclusions if the assumption expressed by the CBI hypothesis does not apply to reality. Second, the CBI hypothesis does actually not suggest a concrete inductive reasoning principle.[7] In fact, in applications it is often used in a more "intuitive" manner. It is hence not astonishing that approaches such as CBR are often criticized for their ad-hoc character. Even though methods of this kind proved to be successful in practice, this criticism might be intelligible from a theoretical point of view. Indeed, it seems to be true that many CBR applications, even if successfully solving the problem at hand, lack a sound theoretical basis.

[5] For example, a cup is similar to a plate as far as its fragility is concerned, but rather different with respect to its functionality.

[6] For example, the relative utility of a good can be expressed in terms of its exchange value.

[7] That is, a principle which justifies inductive conclusions and which allows one to handle such conclusions in a logically consistent way. (Per definitionem, such a principle cannot be a logical tautology.)

Again, it is interesting to make a comparison with other reasoning methods. Inductive statistical inference on the basis of frequency, for example, seems to enjoy a much greater acceptance than similarity-based inference, even though one might argue that it proceeds from assumptions which are hardly less hypothetical. Let us mention, however, an important difference which might explain this situation to some extent: As opposed to similarity-based reasoning, statistical inference is based on (probabilistic) *models* which state the underlying assumptions in an unequivocal and explicit way. Moreover, it makes use of general inference principles, such as MAXIMUM LIKELIHOOD. For such principles it is possible to establish formal properties (e.g. consistency) which do not depend on the specific application. This provides the basis of comparative studies and for estimating the reliability of inference results.

A major concern of this book is to contribute to a theoretical foundation of CBI, thereby improving the understanding, acceptance and applicability of this type of inductive inference. To this end, we shall especially emphasize two aspects:

- In many methods, the CBI hypothesis is used in a more or less implicit way. Thus, the underlying assumptions often remain unclear. To counteract this flaw, we shall propose *explicit models* of the CBI hypothesis. These models can then be taken as a point of departure for developing more transparent similarity-based inference schemes.
- Likewise, the CBI hypothesis and related reasoning principles are often used without testing their validity for the application at hand. Therefore, we are especially interested in improving the *confidence* in CBI, i.e., in making similarity-based inference more reliable. This requires the elaboration of general inference principles and the investigation of their properties. In this connection, the aforementioned models will prove to be useful.

The CBI hypothesis mainly concerns the aspect of *prediction*: It suggests a special approach to exploiting experience in the form of previously observed *cases* (input–output tuples) in order to predict the outcome of a new situation. The term *case-based inference* (CBI) will, therefore, explicitly refer to this performance task, which will be the focus of our investigations. Apart from CBI, which is closely connected with instance-based machine learning methods, similarity-based reasoning might involve additional aspects, of course. In the context of CBR, for instance, CBI does not cover the complete process of problem solving, i.e., the process of ultimately *finding* a solution, but only constitutes the first part thereof.[8] Typically, the characterization of a solution provided via CBI will be utilized by (adaptation) methods applied in subsequent stages of the overall problem solving procedure.

Just as an aside, let us make a remark on the term “prediction”. In statistics, this term often refers to the outcome of a new random sample (given some knowledge

[8] As will be seen later, it basically supports the *retrieval* of relevant cases.

which comes from the observation of other samples generated by the same stochastic process). Besides, it is often assigned a temporal meaning, e.g., in time series analysis. In this book, the term will not especially refer to any of these aspects. Instead, we shall use "prediction" as a generic term for different performance tasks which are concerned with estimating some unknown entity. This includes standard (statistical) estimation problems such as *classification* (assigning an input $x$ to one among a finite set of possible categories), *regression* (estimating or approximating functional relationships with numerical outputs), or *density estimation* (estimating or approximating a stochastic model in the form of a probability density function, given a random sample), but also more general problems involving "complex" output spaces, such as predicting (or rather suggesting) the configuration of a technical system as a candidate solution in case-based problem solving.

### 1.2.1 Making case-based inference more reliable

How can confidence in case-based inference be increased? Of course, the quality of predictions derived on the basis of the CBI hypothesis crucially depends on the validity of this assumption in the context of the application at hand. This validity, in turn, is basically determined by the involved formalization of the similarity concept, i.e., by what one is inclined to call similar (and to which degree).[9]

Thus, one possibility to guarantee correct predictions is simply to choose the "right" measures of similarity. In fact, the CBI hypothesis does trivially apply (and does actually become circular) when using the "ideal" similarity measure, according to which two inputs are similar if the associated outcomes are similar. However, apart from the drawback that this measure of similarity might not be intuitive at all, this approach is clearly not practicable: Since just the outcomes have to be predicted, the "ideal" similarity measure cannot be derived.[10] Even though the quality of predictions might be greatly improved by adapting the similarity measures in a suitable way, it is principally impossible to *guarantee* a good performance by doing so.

In this book, we follow a more pragmatic approach. Roughly speaking, we assume the similarity measures and, hence, a concrete version of the CBI hypothesis to be given. Predictions are then derived on the basis of this particular hypothesis. However, rather than merely suggesting a certain output, we shall pay special attention to the quantification of the *reliability* of a prediction. In fact, not only will it quite often be important to know that a certain output seems possible, but also to have an idea of the degree to which that output is really supported by past experience. From statistical estimation theory, for instance, it is well-known that a point-estimation of a parameter is not worth very much without an associated confidence region, and that the quantification of reliability is critical for

[9] See again the example in footnote 5.

[10] Though it might possibly be estimated from a set of observations.

subsequent decision making. In CBI, this aspect has not received much attention so far.[11]

To illustrate this further, imagine a concrete application and suppose the similar problem–similar solution assumption underlying case-based reasoning to be hardly justified for that application (and related similarity measures). One might then expect an inference result expressing that certain solutions appear rather plausible in connection with a new problem but also that many alternative solutions should not be excluded from the start. Making a prediction of this kind seems more reasonable than simply pointing to one specific solution, thereby disregarding other plausible solutions and pretending a level of certainty which is actually not justified. This becomes especially apparent, e.g., when thinking of a “problem” as a patient with certain symptoms and of the “solution” as a medical diagnosis and a related treatment. In fact, completely disregarding a possible disease or initiating a treatment on the basis of an ill-supported diagnosis might have terrible consequences.

### 1.2.2 The important role of models

In this connection, the advantage of founding predictions (or other types of inference) on explicit models becomes rather obvious. In fact, a model makes the set of assumptions underlying the (inductive) reasoning process transparent. The adequacy of these assumptions can then be verified by confronting the model with observed data, i.e., by performing so-called diagnostic (goodness-of-fit) tests. This, in turn, gives an idea of the confidence of predictions derived from the model. Again, it might be illustrative to compare this with statistical methods. For example, some assumptions on the (data-generating) system under consideration might suggest the specification of a simple linear regression model[12]

$$Y = \alpha + \beta X + \varepsilon. \tag{1.1}$$

This model includes assumptions of different type: structural assumptions concerning the functional dependence between the (random) variables $X$ and $Y$, probabilistic assumptions and assumptions of independence concerning the generation of observed data, and assumptions on the distribution of the error term $\varepsilon$.

While some of the these assumptions are simply taken for granted, others can be tested against the background of observed data. This concerns, e.g., the linear structure of the dependence between $X$ and $Y$. Testing the validity of assumptions

[11] Notable exceptions include the application of general validation procedures such as cross validation or, more generally, methods for estimating the generalization error. Such procedures, however, typically refer to the average performance of a method instead of the validity of individual predictions.

[12] The uncertainty related to a given model is sometimes referred to as “within-model uncertainty.” Often, however, model selection procedures are employed in a first step in order to seek out an optimal model from a class of candidate models. In this case, “between-model uncertainty” must also be taken into account [97].

(in conjunction with properties of the employed inference procedure) in turn allows one to quantify the reliability of estimations (of the parameters $\alpha, \beta$) or predictions (of an output value $y_0$, given a new input $x_0$) made on the basis of model (1.1). For example, such inference results will be unreliable (and endowed with large confidence regions) if the assumption of a linear structure is not correct. In summary, a statistical analysis is generally based on the following information:

I. Assumptions on the system structure (e.g. linear).

II. Assumptions on observed data (e.g. independent and identically distributed).

III. A concrete inference principle (e.g. MAXIMUM LIKELIHOOD).

A major concern of this book is to propose counterparts to points (I) and (III) within the framework of CBI. Thus, our aim is to suggest formal models of the structural assumptions underlying CBI (as expressed by the CBI hypothesis) and, proceeding from such models, to develop related inference procedures.

Let us add that a model is not only good for testing the adequacy of underlying assumptions. Beyond that, it also supports specific adaptation steps which become necessary if these assumptions turn out to be unjustified. In fact, the discussion so far has shown that the validity of the CBI hypothesis largely depends on the application at hand. Thus, a related model has to be adapted to this application in much the same way as a regression model has to be calibrated separately for each set of data. In this connection, we shall emphasize the aspect of model building as well as model adaptation by means of (machine) learning and linguistic modeling techniques.

Finally, a model can also serve as a source of explanation and justification of an inference result. This is an important aspect in knowledge-based systems: In general, a user will be much more satisfied if the system does not only provide the solution itself, but does also give a clue as to how that solution can be justified.

### 1.2.3 Formal models of case-based inference

To realize the above ideas, we shall proceed from a precise interpretation of the CBI assumption: It will be considered as a (deterministic) constraint on degrees of similarity associated with pairs of cases, where the concrete form of the constraint is determined by the system under consideration. This simple model will then be generalized by formalizing the CBI principle within different frameworks of approximate reasoning and reasoning under uncertainty. More precisely, we shall propose *probabilistic* methods as well as models which are based on *fuzzy set* and *possibility theory*. An approach of this kind seems particularly suitable since it emphasizes the heuristic and, hence, uncertain character of CBI. In fact, a *weak version* of the CBI hypothesis according to which similar problems are (at most) *likely* to have similar solutions might be seen as a more appropriate description of this assumption.[13] In the sense of this version, the CBI hypothesis relates the similarity of inputs to the (uncertain) *belief* concerning the (similarity of) outcomes, not to the (similarity of) outcomes directly. Interestingly enough, this approach is to some extent in line with ideas of plausible reasoning introduced by G. Pólya [297]. He complements classical inference patterns such as the *modus ponens*

$$\begin{array}{l} B \text{ implies } A \\ B \text{ is true} \\ \hline A \text{ is true} \end{array}$$

by different schemes of *plausible inference.* A typical example, in which the concept of similarity (analogy) is directly related to that of *plausibility*, is the following inference pattern:

$$\begin{array}{l} A \text{ is analogous to } B \\ B \text{ is true} \\ \hline A \text{ is more plausible} \end{array}$$

Once its underlying hypothesis has been formalized within a certain framework of reasoning under uncertainty, CBI can benefit from related inference procedures. A formalization in the framework of possibility theory, for instance, allows one to exploit (fuzzy set-based) approximate reasoning techniques in order to realize similarity-based inference. Likewise, a probabilistic model makes the powerful methodological framework of statistical inference accessible to CBI.[14] This way, it becomes possible to equip CBI with a solid basis which is grounded on well-established reasoning and inference techniques.

The different formalizations of CBI introduced in subsequent chapters should not be seen as competing methods. Rather, they emphasize different aspects. The probabilistic and the possibilistic[15] approach, for instance, complement each other in a reasonable way. In fact, the superiority of a particular model depends strongly on the respective application. Let us only mention a rough distinction which is related to the available information sources and the problem of learning and knowledge acquisition: In general, the probabilistic approach leads to methods which can be qualified as data-driven. These methods turn out to be very efficient when disposing of a large sample of observed cases and might hence be preferred on such grounds. By making use of fuzzy set-based modeling techniques, possibilistic models are particularly suitable for incorporating domain-specific (expert) knowledge into CBI models. In fact, they provide a convenient

[13] Interestingly enough, the probabilistic nature of inductive generalization has been emphasized by philosophically minded scholars for a long time [266].

[14] As will be seen in subsequent sections, not only can CBI benefit from statistical methodology but also vice versa.

[15] In analogy to the term "probabilistic", we employ the neologism "possibilistic" in the sense of "based on possibility theory".

basis for combining knowledge-driven and data-driven reasoning. Consequently, they might be preferred if data is sparse or of low quality and if this lack of information can be compensated by available background knowledge.

## 1.3 Overview

In the first part of the book, the aforementioned models of the CBI hypothesis and the related reasoning principles are developed. The second part is devoted to applications in the fields of *decision making* and *problem solving*. Here, we give a brief outline of the chapters that follow.

**Similarity and case-based inference**

This chapter provides some background information on case-based (instance-based) reasoning and on the formalization of the similarity concept. We briefly survey some methods in which similarity plays a central role, notably NEAREST NEIGHBOR classification, instance-based learning, and case-based reasoning. Moreover, we introduce a formal framework of *case-based inference* (CBI) which provides a common basis for the methods proposed in subsequent chapters.

**Constraint-based modeling of case-based inference**

In this chapter, we introduce a basic model of CBI. To this end, we proceed from a constraint-based interpretation of the CBI hypothesis, according to which the similarity of inputs imposes a constraint on the similarity of associated outcomes in the form of a lower bound. The concept of a *similarity hypothesis* is introduced as a formalization of this interpretation.

In connection with a set of observations, a similarity hypothesis allows for realizing case-based inference as a kind of constraint propagation. This inference scheme leads to *set-valued* predictions, i.e., an unknown output is characterized by a set of possible candidates. We propose an efficient algorithm for learning similarity hypotheses from observed data. By making use of the hypotheses thus derived, one obtains set-valued predictions which are as precise as possible. At the same time, these predictions are *probably correct*: A set of predicted candidates covers the true outcome with high probability (and is hence referred to as a "credible output set"). Not only are such properties derived analytically, they are also validated by means of experimental studies.

Finally, we outline some applications of (constraint-based) CBI in the context of *statistical inference*, namely similarity-based parameter estimation and the similarity-based elicitation of priors in Bayesian analysis. In this connection, CBI is applied as a non-parametric approach to estimating confidence regions.

### Probabilistic modeling of case-based inference

In this chapter, we propose a generalization of the constraint-based approach in which the similarity of inputs is taken as an indication of the *probability* that the associated outputs are similar (to a certain degree). A probabilistic formalization of this kind is more flexible and permits for "exceptions to the (CBI) rule."

We introduce a basic probabilistic model of the CBI hypothesis, called a *probabilistic similarity profile*. Taking this model as a point of departure, different types of case-based inference schemes are realized. In particular, we propose a model in which cases are considered as individual pieces of (uncertain) evidence, where evidence is represented in the form of *belief functions*. This idea is formalized within a framework of *information fusion*. In this context, we also discuss the assessment of cases based on the reliability and precision of the estimations they provide. Related to this aspect is the idea of an "exceptional case" and the problem of discovering and discounting such cases. Moreover, we develop an alternative CBI scheme that derives approximate inference results in the form of upper probability bounds.

In connection with the learning of similarity measures and hypotheses, this chapter establishes some interesting relations between case-based reasoning and statistical methods. It is argued that the two fields can cross-fertilize each other: On the one hand, a probabilistic formalization of case-based reasoning makes the powerful methodological framework of statistics accessible to CBI. On the other hand, the idea of reasoning on the basis of similarity appears interesting from the viewpoint of statistical modeling as well and might contribute to the extension of existing statistical methods.

### Fuzzy set-based modeling of case-based inference

In Chapters 5 and 6, a formal framework of case-based inference is presented in which the generalization beyond experience is founded on the concepts of similarity and *possibility*. The underlying extrapolation principle is formalized within the framework of *fuzzy rules*. Thus, case-based reasoning can be realized as fuzzy set-based approximate reasoning. Fuzzy rules establish a relation between the concepts of similarity and possibility which takes the uncertain character of case-based inference into account: Extrapolation is "possibilistic" in the sense that predictions take the form of possibility distributions on the set of outcomes, rather than precise (deterministic) estimations.[16] Moreover, the aspect of confidence in CBI is again taken into account: The generalization of case-based information is founded on a model (of the CBI principle) in the form of a set of fuzzy rules. That is, a prediction is always justified by an

[16] Since a set (resp. its characteristic function) can be identified with a special ({0,1}-valued) possibility measure, this type of prediction can again be seen as a generalization of set-valued predictions as derived in connection with the constraint-based approach.

explicit rule which allows one to conclude from the similarity of inputs on the possibility of certain outcomes.

Moreover, the close connection between possibility theory and fuzzy sets allows for exploiting the merits of linguistic modeling techniques in the context of CBI. In fact, linguistic modeling provides a convenient way of incorporating domain-specific (expert) knowledge. The approach thus allows one to combine knowledge and data in a flexible way and favors a view of CBI according to which the user interacts closely with the reasoning system.

In Chapter 5, the CBI hypothesis is formalized by means of so-called *possibility rules*, an example-based type of fuzzy rule that has originally been applied in the context of fuzzy control. The basic principle of the related inference mechanism is a similarity-guided extrapolation of observed cases. According to this principle, an encountered case is taken as evidence for the existence of similar cases. This evidence is expressed in terms of degrees of possibility assigned to hypothetical cases and thus defines a possibilistic approximation of an underlying (but only partially observed) set of existing cases. Expert knowledge can be incorporated into the model by means of (linguistic) modifier functions acting on rules and related similarity measures. Besides, these functions provide the basis of calibration (learning) methods which adapt the model to the application at hand. Here, the idea is to proceed from a purely linguistic specification of a CBI model. A concrete model is then obtained by adapting the broad structure thus defined to the observed data.

In Chapter 6, we make use of *implication-based* fuzzy rules which involve a complementary principle of case-based reasoning. Roughly speaking, such rules realize a *constraint-based* approach[17] in the sense that encountered cases are considered as evidence for (partially) *excluding* certain other (hypothetical) cases, which are *not similar* enough to the observed ones. This is to be contrasted with conjunction-based rules, where memorized cases are considered as pieces of data which *support* the possibility of observing *similar* cases.

### Case-based decision making

This chapter deals with the idea of applying the CBI principle in the context of decision making: An agent faced with a decision problem relies upon its experience from similar problems in the past. That is, it chooses an act based on the performance of (potential) acts in preceding problems which are similar to the current one. A formal framework of case-based decision making, intended as an alternative to expected utility theory, has originally been introduced by Gilboa and Schmeidler [167]. Beyond that, Dubois and Prade [101] have outlined a related formalization in the context of (fuzzy set-based) approximate reasoning. We briefly review these methods and propose some extensions and generalizations thereof. Moreover, we develop a new approach based on

[17] In fact, this approach can be seen as a generalization of constraint-based CBI as realized in Chapter 3.

methods of case-based inference as introduced in previous chapters. This approach basically differs from previous proposals in that observed cases are not used for selecting an act directly. Rather, such cases have influence on the decision maker's uncertainty about the consequence of choosing a certain act. It leads to an extended and more expressive decision-theoretic setup which combines the (cognitive) concepts of belief, preference and similarity.

### Case-based problem solving

This chapter is devoted to applications of similarity-based inference and statistical methods in the context of heuristic search. We discuss and formalize the idea of repetitive (search-based) problem solving, understood as the repeated use of heuristic search for solving problems which share a similar structure. For obvious reasons, this type of problem is particularly interesting from the perspective of CBI and statistics. The basic idea is to exploit the experience from (combinatorial optimization) problems already solved in order to improve the efficiency of future (search-based) problem solving. The chapter contributes from various directions to a methodological framework in which (repetitive) problem solving by heuristic search can be realized: Firstly, it elaborates on the idea of making search more efficient by using case-based inference in order to *focus the search process* on promising regions of the search space. Secondly, a novel search strategy based on statistical methods of *change detection* is introduced. Thirdly, we propose a statistical method for estimating evaluation functions which control the choice of search operators. Such functions are used in order to guide a heuristic search algorithm for solving resource-based configuration problems, a particular application which is closely related to integer linear programming.

# 2. Similarity and Case-Based Inference

This chapter serves two purposes. Firstly, we provide some background information on similarity-based reasoning and related topics. Secondly, we introduce a formal framework of *case-based inference* (CBI) that provides the basis for the methods which are developed in subsequent chapters.

Section 2.2 briefly reviews some problem solving methods in which the concept of similarity (or the dual concept of distance) plays a central role. From a machine learning (statistical) perspective, these methods generally belong to the so-called *instance-based* (non-parametric) approaches. Therefore, we begin with a brief comparison of *model-based* and instance-based reasoning in Section 2.1. The concept of similarity itself and formalizations thereof are discussed in Section 2.3. In that section, we shall also outline a new approach to similarity evaluation which makes use of so-called fuzzy integrals as aggregation operators. In Section 2.4, we introduce the aforementioned framework of CBI.

## 2.1 Model-based and instance-based approaches

The major concern of disciplines such as machine learning and statistical inference is to generalize beyond observed data.[1] To this end, the observations encountered are interpreted against available background knowledge. This knowledge, often not more than a set of assumptions, is generally expressed (or rather encoded) in terms of a *hypothesis space* $\mathcal{H}$, such as a certain class of rule bases in rule induction or linear functions in regression analysis.[2] The hypothesis space, in conjunction with a related strategy for searching this space, determine (at least partially) what is called the *inductive bias* in machine learning. Loosely speaking, the inductive bias is responsible for choosing among the possible (consistent or equally acceptable) generalizations. More specifically, a bias reducing the class of admissible hypotheses (that is, the hypothesis space) is called a *restriction bias* or *language bias*. As opposed to this, a *search bias* or *preference bias* has influence on the generalization through the way a hypothesis space is searched: Using a certain search strategy, one acceptable hypothesis might be found before and,

[1] Being aware of the grave philosophical and logical objections to this venture (as raised by HUME and still not solved in a satisfactory way, neither by POPPER's deductivism nor by CARNAP's inductivism).

[2] In some machine learning methods, e.g., explanation-based learning or inductive logic programming, background knowledge is represented in a more explicit way.

hence, given priority over another one (the latter is left out of account if the first hypothesis is immediately adopted).

Since the observed data is usually consistent with a multitude of generalizations, it would in fact be meaningless without a "biased" angle of view [270]. A strong bias generally comes along with a restrictive set of prior assumptions. Thus, it leads to a small space of (still admissible) hypotheses, a prerequisite for good generalization performance. As opposed to this, an extremely rich (complex) class of hypotheses[3] (models) is often capable of a very accurate adaptation to the observed data, but performs weak when generalizing beyond these data.[4]

### 2.1.1 Model-based approaches

A model (hypothesis) $h^* \in \mathcal{H}$ is chosen on the basis of how well it fits the observed data according to some criterion, e.g., as expressed by the Maximum Likelihood principle. Further aspects such as the simplicity of the model might also be considered, especially if several models fit the data equally well. The principle of Occam's Razor[5] [37], for instance, suggests to prefer the simplest hypothesis that fits the data.[6] Once having selected a specific model $h^*$, it can be used for various types of performance tasks, such as explanation (of the observed data), problem solving, and prediction (of future observations). Given a new input $x$, for instance, the associated output might be estimated by $\widehat{y} = h^*(x)$.

In general, inductive inference performing along these lines is realized by *parametric* statistical methods and *model-based* machine learning techniques. Such methods aim at constructing a low-dimensional (parametric) model which – at least to some degree – explains the observed data. They often make explicit assumptions about some underlying data-generating process, e.g., in the form of relational dependencies between observable quantities and statistical distributions of random variables.

It is important to distinguish different types of models according to which aspects of the modeled system are taken into consideration. A simple equation relating some variables (as in linear regression) might be sufficient for making predictions (and for acting accordingly), but it will generally not allow for understanding a system or for explaining the mechanism underlying the observed phenomena. A corresponding distinction between "interpolatory formulae" and "explanatory

[3] It is actually not the size of the hypothesis space which determines complexity and hence generalization performance. A useful measure of the complexity of a hypothesis space is the Vapnik-Chervonenkis (VC) dimension [379].

[4] The choice of an adequate hypothesis space is an important problem in machine learning, and some attempts at automating this choice have already been made (e.g. [24]). Corresponding ideas are also discussed under the slogan "learning to learn" [368].

[5] An alternative (maybe more correct) spelling of the well-known philosopher's name is "Ockham".

[6] Of course, one still has to specify the meaning of simplicity in order to make this principle applicable. For instance, the Minimum Description Length principle can be used for putting Occam's Razor into practice. See [91] for a critical investigation of Occam's Razor.

models" has already been made by Neyman [279]. A similar differentiation between "surface models" and "deep models" has recently given rise to the founding of *model-based reasoning*, which is now a proper subfield of AI. Needless to say, arbitrary gradations exist between these extreme types of models [248]. Subsequently, we shall use the term "model-based" (learning) merely as a counterpart to "instance-based" (learning).

### 2.1.2 Instance-based approaches

Instance-based methods represent an alternative approach to (machine) learning. In model-based learning, an observation has an indirect and global influence on predictions in the sense that it generally affects the complete set of parameters which specify the model. Predictions are then derived from that model. As opposed to this, individual observations contribute in a more direct but often locally limited way to the inference result in instance-based methods. A typical example is classification according to the Nearest Neighbor (NN) principle (cf. Section 2.2).

Instance-based approaches generally belong to the class of so-called *lazy* learning methods [6] (also known as memory-based [359], exemplar-based [327] or case-based [234]). These methods learn by simply storing (some of) the observed examples. They defer the processing of these inputs until a prediction (or some other type of query) is actually requested. Predictions are then derived by somehow combining the stored examples. After the query has been answered, the prediction itself and any intermediate results are discarded. Again, the NN classifier is a typical example of a lazy (instance-based) learning method. Locally weighted regression is an example of a (statistical) method which is lazy in the same sense. In fact, instance-based approaches to machine learning share important features with *non-parametric* methods in statistics, such as kernel smoothing techniques [385]. It deserves mentioning, however, that instance-based methods are not necessarily non-parametric.[7]

As opposed to lazy learners, model-based methods are *eager* in the sense that they greedily compile their inputs into an intensional description (model), such as a decision tree or a regression function, and then discard the inputs. The induced description is used in order to reply to future information requests. Model-based learning is thus in line with parametric methods in (classical) statistics.

### 2.1.3 Knowledge representation

One might argue that instance-based reasoning avoids (or at least circumvents) the philosophical problem of induction. This is why instance-based reasoning is

[7] Even the Nearest Neighbor classifier can be seen as an instantiation of a parameterized(!) class of (lazy) learning algorithms [396].

called *implicit induction* in [172], where it is contrasted with (explicit) induction as realized, e.g., when learning general rules from specific examples. Indeed, the lazy learning paradigm is naturally related to what is called *transductive inference* in statistical learning theory [380]. Transductive inference is inference "from specific to specific." Thus, it stands for the problem of estimating some values of an unknown functional relation $f(\cdot)$ *directly*, given a set of empirical data. Roughly speaking, the test set of points for which predictions have to made later on is already known in the training phase. This type of inference represents an alternative to the indirect (model-based) approach which derives an estimation $h^*$ of the *complete* functional relationship in a first step (induction) and evaluates this estimation at the points of interest afterwards (deduction).

In connection with lazy (instance-based) learning, we shall also speak about "extrapolation" of (the information coming from) observed cases. In fact, typically the known values of the function $f(\cdot)$ are extrapolated – in a locally limited way – in order to estimate unknown values. Of course, by making predictions both, instance-based as well as model-based methods do generalize beyond observed data. They only employ different types of knowledge representation: extensional descriptions in instance-based reasoning and intensional descriptions in model-based methods. Consequently, the generalization step is delayed in instance-based methods where it corresponds to the extrapolation of observed cases. As opposed to this, generalization corresponds to induction (model building) in model-based approaches. In this connection, it is important to note that an extensional description in the form of a set of observed cases is nothing else than a collection of known facts. Strictly speaking, it can never be false,[8] whereas an intensional description induced from the observed data might be incorrect.

Instance-based methods still have to incorporate some kind of background knowledge into their process of generalization, although they do not refer to an explicit model. This inductive bias generally corresponds to some sort of representativeness or closeness assumption. As already mentioned in Chapter 1, this assumption is most clearly expressed by the CBI hypothesis, suggesting that "similar causes bring about similar effects." Besides, the application of the CBI principle assumes a precise idea of the concept of similarity in the respective context, i.e., it makes the quantification of similarity necessary. Since the definition of similarity measures will again fall back on some underlying assumptions, one can indeed question the existence of something like genuine model-free reasoning [235].

### 2.1.4 Performance in generalization

Of course, comparing the quality of predictions is possible only for specific methods, not for instance-based and model-based approaches in general. But even

[8] At least when disregarding philosophical tenets which do even challenge the possibility of knowing facts.

for specific methods it is often difficult to make a comparison since performance depends strongly on the application and properties of the data.[9]

Still, it can roughly be said that model-based approaches are in general more knowledge-oriented, since constructing a model often involves a considerable amount of background knowledge. The observed data is then merely used for calibrating this knowledge. Consequently, a model-based approach will yield good inference results if the model is suitably defined, but it might lead to erroneous conclusions if the assumed model structure does actually not apply to reality. Instance-based methods are more data-oriented and present a way to overcome this danger.[10] Nevertheless, the lack of predefined structure can lead to problems such as the *overfitting* of data. Let us mention that several semi-parametric statistical methods have been proposed in order to combine the merits of both, parametric and non-parametric approaches (at the cost of an often increased computational complexity). The same idea has given rise to the emergence of hybrid (integrated) approaches in machine learning [307].

### 2.1.5 Computational complexity

Learning in instance-based methods is rather simple from a conceptual point of view. Basically, it amounts to storing new experiences in the form of observed cases. Needless so say, however, simply adding all observations to the memory is generally not the best strategy. In fact, the larger the number of stored cases, the larger the time complexity and memory requirements of the inference procedure. In order to optimize performance one has to use a more sophisticated strategy of maintaining a memory of cases, which does also allow to remove already stored cases (cf. Section 2.2). If it is possible to influence the choice of the next query (problem), one might even try to control this choice so as to complement the current experience in the most reasonable way.

Briefly, learning in instance-based methods can be seen as organizing an optimal memory of cases.[11] This is often simpler and more efficient than learning in model-based approaches. Particularly, instance-based learning is inherently *incremental*. In fact, an extensional description can be updated quite easily, namely by adding a new case (fact). As opposed to this, the adaptation of an induced model which becomes necessary due to the observation of a new case is often much more difficult. In non-incremental induction, adaptation comes down to re-estimation, i.e., to deriving a new model from scratch.[12] As can be seen, from an estimation point of view the problem with non-incremental model-based approaches is not

[9] The observation that each approach to (inductive) learning works best in some special domain, but not in general, has been termed the *selective superiority problem* in machine learning [57].

[10] Still, they generally require the definition of a reasonable similarity or distance measure.

[11] Of course, here we neglect other aspects of learning, such as adapting the similarity measure.

[12] Note that this requires to store not only the model, but also the complete data.

the one of induction (as discussed above), but rather the related one of knowledge revision.[13]

The derivation of predictions is generally efficient in model-based approaches, where it often comes down to evaluating a functional expression. Compared to this, the combination of stored cases in lazy methods is less efficient (among other things, it requires the searching of the memory).

From a complexity point of view, it can hence be said that lazy (instance-based) methods have lower computational costs than eager (model-based) algorithms during the training phase. Moreover, knowledge revision can be realized in a simple and straightforward way. On the other hand, they generally have greater storage requirements (typically linear in the size of the data set) and higher computational costs when it comes to deriving a prediction.

## 2.2 Similarity-based methods

### 2.2.1 Nearest neighbor (NN) estimation

The well-known NEAREST NEIGHBOR (NN) principle, which originated in the field of pattern recognition [76], provides an intuitively simple approach to the prediction of both categorical and numerical outputs.

To introduce and discuss the main ideas of NN estimation, consider an input space $\mathcal{X}$ endowed with a distance measure $\Delta_{\mathcal{X}}$.[14] The elements of $\mathcal{X}$ are *instances* $x$ which can be though of as the description of objects (usually in attribute–value form). $\mathcal{L}$ denotes a set of outputs, and $\langle x, \lambda_x \rangle \in \mathcal{X} \times \mathcal{L}$ is called a labeled instance (or a case). In classification tasks, $\mathcal{L}$ is a finite (usually small) set $\{\lambda_1, \ldots, \lambda_m\}$ comprised of $m$ class labels. Let $S$ denote a sample that consists of $n$ labeled instances $\langle x_\imath, \lambda_{x_\imath} \rangle$ $(1 \leq \imath \leq n)$. Finally, a new instance $x_0 \in \mathcal{X}$ is given, whose label $\lambda_{x_0}$ is to be estimated.

With regard to the sample $S$, note that $\mathcal{X} \times \mathcal{L}$ corresponds to the set of *potential* observations. For each label $\lambda \in \mathcal{L}$, let $C_\lambda \subseteq \mathcal{X}$ denote the set of instances $x \in \mathcal{X}$ such that $\langle x, \lambda \rangle$ can indeed be observed. $C_\lambda$ is also referred to as a *concept*. Formally, one can assume an underlying population of entities such that each element $e$ of this population is mapped to a labeled instance $\langle x(e), \lambda(e) \rangle$ in a unique way. Thus, $x$ is an element of $C_\lambda$ or, say, $\langle x, \lambda \rangle$ is an *existing* instance if there is at least one $e$ such that $\langle x, \lambda \rangle = \langle x(e), \lambda(e) \rangle$. Note that the mapping $e \mapsto x(e)$ is not assumed to be injective (different elements of the population might have the same description), which means that concepts can overlap ($C_\lambda \cap C_{\lambda'} \neq \emptyset$ for $\lambda \neq \lambda'$).

[13] Yet, it should be recognized that model-based approaches are not necessarily non-incremental.

[14] $(\mathcal{X}, \Delta_{\mathcal{X}})$ is often supposed to be a metric space.

The NN principle prescribes to estimate the label of the yet unclassified instance $x_0$ by the label of the closest sample instance, viz the one that minimizes the distance to $x_0$. The $k$-Nearest Neighbor ($k$NN) approach is a slight generalization which takes the $k > 1$ nearest neighbors of a new query $x_0$ into account. That is, an estimation $\lambda_{x_0}^{est}$ of $\lambda_{x_0}$ is derived from the set $\mathcal{N}_k(x_0)$ of the $k$ nearest neighbors of $x_0$, e.g., by means of the *majority vote* decision rule:

$$\lambda_{x_0}^{est} = \arg\max_{\lambda \in \mathcal{L}} \operatorname{card}\{x \in \mathcal{N}_k(x_0) \mid \lambda_x = \lambda\}. \tag{2.1}$$

Not only can the NN principle be used for classification, it is also employable for realizing a (locally weighted) approximation of continuous-valued target functions. To this end, one reasonably computes the (weighted) mean of the $k$ nearest neighbors of a new query point instead of returning the most common value.[15]

The inductive bias underlying the NN principle corresponds to a *representativeness* or *closeness* assumption suggesting that similar (= closely located) instances have similar (or even the same) label. This assumption is obviously a special version of the CBI hypothesis (cf. Section 1.1). It gives rise to a similarity-guided extrapolation principle which is clearly of a heuristic nature. Still, theoretical properties of NN classification have been investigated thoroughly from a statistical perspective (e.g. [74]).[16] In fact, the origin of the NN approach can be found in work on non-parametric discriminatory analysis [148, 149].

Besides, several conceptual modifications and extensions, such as distance weighting, which is discussed below, have been considered. Particularly, (editing) methods for selecting optimal training samples to be stored in the memory have been developed in order to improve classification performance [163, 397] or to reduce computational complexity [186] or both. Other extensions aim at supporting the determination of adequate metrics [392] and the optimal size of the neighborhood. Computational aspects have been addressed as well. For example, fast algorithms and efficient data structures for finding nearest neighbors have been devised in order to improve computational efficiency [154, 158, 411, 223, 222, 287].

**Uncertainty in NN estimation.** In statistical estimation theory, an estimated quantity is always endowed with a characterization of its reliability, usually in terms of a confidence measure and a confidence region. Alternatively, an estimation is given directly in the form of a probability distribution. As opposed to this, the NN principle in its basic form merely provides a point-estimation or, say, a decision rule, but not an estimation in a statistical sense. The neglecting of uncertainty makes this principle appear questionable in some situations, a point that we shall return to in later chapters. To illustrate, Fig. 2.1 shows two classification problems. The new instance $x_0$ is represented by a cross, and dark and light circles correspond to instances of two different classes, respectively. In both

[15] Shephard's interpolation method [340] can be considered as a special type of NN estimation.

[16] Needless to say, corresponding results can only be derived under certain statistical assumptions on the setting of the problem.

cases, the $k$NN rule with $k = 5$ suggests DARK as a label for $x_0$. As can be seen, however, this classification is everything but reliable: In the above setting, the proportion of dark and light examples is almost balanced (apart from that, the closest points are light). This is a situation of *ambiguity.* The setting below illustrates a problem of *ignorance*: It is true that all neighbors are dark, but even the closest among them are actually quite distant.

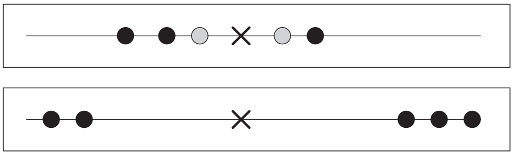


**Fig. 2.1.** Two situations of uncertainty in connection with the basic $k$NN rule, caused by the existence of more than one frequent class label among the nearest neighbors (above) and the absence of any close neighbor (below).

A simple (yet drastic) step to handle this type of problem is to apply a reject option in the form of a distance or frequency threshold. That is, a classification or answer to a query is simply refused if the nearest neighbors are actually not close enough [370, 75, 136] or if the most frequent label among these neighbors is still not frequent enough [68, 188].

A second possibility is to equal statistical methods (especially Bayesian ones) in deriving a probability distribution as an inference result. In fact, this is an obvious idea since NN techniques have originally been employed in the context of non-parametric density estimation [148, 256]. Thus, a single decision can be replaced by an estimation in the form of a probability vector

$$\big(p_{x_0}(\lambda_1), \ldots , p_{x_0}(\lambda_m)\big) \, , \tag{2.2}$$

where $p_{x_0}(\lambda_\imath) = \mathbb{P}(\lambda_\imath \,|\, x_0)$ is the probability that $\lambda_{x_0} = \lambda_\imath$, i.e., the conditional probability of the label $\lambda_\imath$ given the instance $x_0$. Taking the $k$ nearest neighbors of $x_0$ as a point of departure, an intuitively reasonable approach is to specify the probability $p_{x_0}(\lambda_\imath)$ by the relative frequency of the label $\lambda_\imath$ among the labels of these neighbors: $p_{x_0}(\lambda_\imath) \overset{\mathrm{df}}{=} k_\imath/k$, where $k_\imath$ denotes the number of neighbors having label $\lambda_\imath$. In fact, this approach can also be justified theoretically, as will be shown in the following.

The Nearest Neighbor approach to *density estimation* (not to be confused with the one to classification) is closely related to kernel-based density estimation. An NN density estimator is a kernel estimator with variable kernel width [343]: The size of the neighborhood of a point $x_0$ is adapted so as to include exactly $k$ observations. Thus, consider a sample of $n$ observations $x_1, \ldots , x_n \in \mathfrak{R}^l$ which are realizations of an $l$-dimensional random vector $X$ with probability density $\phi : \mathfrak{R}^l \longrightarrow \mathfrak{R}_{\geq 0}$. For $x_0 \in \mathfrak{R}^l$ let $v$ be the volume of the smallest sphere $V(x_0)$ around $x_0$ that contains $k$ of these observations. The relation

$$\mathbb{P}(X \in V(x_0)) \approx \phi(x_0) \cdot v$$

(which holds true for small spheres) then suggests the following estimation of $\phi(x_0)$, the density at point $x_0$:

$$\phi^{est}(x_0) = \frac{k}{n \cdot v} \tag{2.3}$$

Coming back to NN classification, consider a sample $S$ that comprises $n = n_1 + \ldots + n_m$ observations, where $n_\imath$ denotes the number of tuples $\langle x, \lambda_x \rangle \in S$ such that $\lambda_x = \lambda_\imath$. Let $x_0$ be a new observation. Again, we choose an as small as possible hypersphere around $x_0$ which contains a set $\mathcal{N}_k(x_0)$ of $k$ instances from $S$, where $k = k_1 + \ldots + k_m$ with $k_\imath = \operatorname{card}\{x \in \mathcal{N}_k(x_0) \,|\, \lambda_x = \lambda_\imath\}$. The conditional probability density of $x_0$ (given the label) can now be estimated by

$$\phi^{est}(x_0 \,|\, \lambda_\imath) = \frac{k_\imath}{n_\imath \cdot v}, \tag{2.4}$$

where $v$ denotes the volume of the hypersphere around $x_0$. Moreover, the unconditional density of $x_0$ and the prior probability of the label $\lambda_\imath$ can be estimated by

$$\phi^{est}(x_0) = \frac{k}{n \cdot v}, \quad p^{est}(\lambda_\imath) = \frac{n_\imath}{n}, \tag{2.5}$$

respectively. For the probabilities in (2.2) one thus obtains

$$p_{x_0}(\lambda_\imath) = p^{est}(\lambda_\imath \,|\, x_0) = \frac{\phi^{est}(x_0 \,|\, \lambda_\imath) \cdot p^{est}(\lambda_\imath)}{\phi^{est}(x_0)} = \frac{k_\imath}{k}. \tag{2.6}$$

Remark 2.1. Note that the NN estimation of the conditional probability density (2.4) is actually given by

$$\phi^{est}(x_0 \,|\, \lambda_\imath) = \frac{k_\imath}{n_\imath \cdot v_\imath},$$

where $v_\imath$ is the volume of the smallest sphere around $x_0$ that contains all of the $k_\imath$ neighbors with label $\lambda_\imath$. Then, however, the probabilities

$$p_{x_0}(\lambda_\imath) = \frac{k_\imath \cdot v}{k \cdot v_\imath} \tag{2.7}$$

do not necessarily add up to 1. This problem is related to a general difficulty of NN density estimation. Namely, deriving (2.3) for all $x \in X$ leads to a non-normalized density function $\phi^{est}$ since each $x$ requires a different hypersphere.[17] □

Of course, (2.6) might be considered as a formal justification of the original $k$NN (decision) rule: The label estimated by the (majority vote) $k$NN rule is just the one of maximal (posterior) probability [79]. Still, one should be cautious

[17] Apart from that, an NN density estimation may suffer from very heavy tails and an infinite integral.

with the distribution (2.6). Particularly, it is not clear how reliable the estimated probabilities $p_{x_0}(\lambda_\imath) = k_\imath / k$ actually are. It is possible to construct corresponding confidence intervals, but these are only asymptotically valid [343]. In fact, $k$ is generally small and, hence, (2.6) not very reliable.[18] Improving the quality of predictions by simply increasing $k$ obviously does not work since it also entails an enlarging of the hypersphere around $x_0$.[19]

**Weighted NN rules.** A straightforward modification of the $k$NN rule is to weight the influence of a neighboring sample point by its distance. This idea leads to replace (2.1) by

$$\lambda_{x_0}^{est} = \arg\max_{\lambda \in \mathcal{L}} \sum_{x \in \mathcal{N}_k(x_0)\,:\,\lambda_x = \lambda} \omega(x \,|\, x_0, S), \tag{2.8}$$

where $\omega(x \,|\, x_0, S)$ is the weight of the neighbor $x$. There are different possibilities to define these weights. For example, let the neighbors $\mathcal{N}_k(x_0) = \{x_1, \ldots, x_k\}$ be arranged such that $d_\imath = \Delta_{\mathcal{X}}(x_\imath, x_0) \leq \Delta_{\mathcal{X}}(x_\jmath, x_0) = d_\jmath$ for $\imath \leq \jmath$. In [137], the weights are then determined as[20]

$$\omega(x_\imath \,|\, x_0, S) = \begin{cases} \frac{d_k - d_\imath}{d_k - d_1} & \text{if} \quad d_k \neq d_1 \\ 1 & \text{if} \quad d_k = d_1 \end{cases} . \tag{2.9}$$

The weighting of neighbors appears reasonable from an intuitive point of view. For instance, a weighted $k$NN rule is likely to yield LIGHT rather than DARK as a classification in Fig. 2.1 (above). More general evidence for the usefulness of distance-weighting is provided in [257, 288], at least in the practically relevant case of finite samples. In fact, in [20] it was shown that the *asymptotic performance* of the $k$NN rule is not improved by distance-weighting.

Note that the original $k$NN rule corresponds to the weighted rule with

$$\omega(x \,|\, x_0, S) = \begin{cases} 1 & \text{if} \quad x \in \mathcal{N}_k(x_0) \\ 0 & \text{if} \quad x \notin \mathcal{N}_k(x_0) \end{cases} . \tag{2.10}$$

Thus, the NN rule can be expressed as a global principle involving the complete sample $S$ of observations without loss of generality:

$$\lambda_{x_0}^{est} = \arg\max_{\lambda \in \mathcal{L}} \sum_{\langle x, \lambda_x \rangle \in S\,:\,\lambda_x = \lambda} \omega(x \,|\, x_0, S). \tag{2.11}$$

Interestingly enough, it is also possible to consider the probabilistic NN prediction (2.6) in the context of the weighted NN approach. Namely, (2.6) can be written as

[18] An estimated probability is always a multiplicity of $1/k$. Particularly, $p_{x_0}(\lambda_\imath) \in \{0, 1\}$ in the special case $k = 1$, i.e., for the 1NN rule.

[19] Good estimations are obtained for *small* hyperspheres containing *many* points. Besides, asymptotic convergence generally assumes an adaptation of $k$ as a function of $n$.

[20] See [257] for a modification that performed better in experimental studies; for other types of weight functions see, e.g., [398].

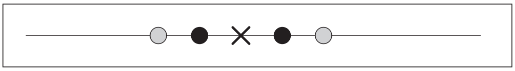
$$p_{x_0}(\lambda) = \sum_{\langle x, \lambda_x \rangle \in S \,:\, \lambda_x = \lambda} \omega(x \,|\, x_0, S), \tag{2.12}$$

with the weight function $\omega$ now being defined by

$$\omega(x \,|\, x_0, S) = \begin{cases} 1/k & \text{if} \quad x \in \mathcal{N}_k(x_0) \\ 0 & \text{if} \quad x \notin \mathcal{N}_k(x_0) \end{cases} . \tag{2.13}$$

Again, (2.11) then amounts to choosing the label with maximal posterior probability.

Of course, in the following situation one would hardly advocate a uniform distribution suggesting that labels DARK and LIGHT have the same probability:



This example reveals a shortcoming of the weight function (2.13), namely the disregard of the *arrangement* of the neighbors [198]. In fact, the derivation of the probabilistic NN estimation (2.6) disregards the actual distances and positions in the estimation of probability densities.[21] This, however, is only justified if the sphere containing the $k$ nearest neighbors is indeed very small, which is usually not the case in practice. (Note that the label DARK is assigned a higher degree of probability than LIGHT according to (2.7), cf. Remark 2.1).

In order to account for this problem, it is possible to combine the idea of weighting and probabilistic estimation. The use of the uniform weights (2.13) corresponds to the use of the (uniform) Parzen window in kernel-based density estimation [289]. By making use of a more general kernel function $K : \mathfrak{R}^l \longrightarrow \mathfrak{R}_{\geq 0}$, a density function which is usually symmetric around 0, the NN density estimation (2.3) can be generalized as follows:

$$\phi^{est}(x_0) = \frac{1}{n} \cdot \sum_{\imath=1}^{n} K_{d_k}(x_0 - x_\imath), \tag{2.14}$$

where $d_k$ is the distance between $x_0$ and its $k$-th nearest neighbor and $K_{d_k}$ is a re-scaling of a kernel function $K$ (with $K(u) = 0$ for $|u| > 1$):

$$K_d : u \mapsto 1/d^l \cdot K(u/d).$$

The same reasoning as in Section 2.2.1 then suggests a weighted counterpart of (2.6):

$$p^{est}(\lambda \,|\, x_0) \propto \sum_{\langle x, \lambda_x \rangle \in S \,:\, \lambda_x = \lambda} K_{d_k}(x_0 - x). \tag{2.15}$$

[21] Taking positions into account becomes very tricky in instance spaces of higher dimension [420].

As can be seen, (2.15) is nothing else than an estimation derived from the weighted NN rule by means of normalization.[22] Thus, proceeding from weights such as (2.9), one simply defines a probability distribution $p_{x_0}$ such that

$$p_{x_0}(\lambda) \propto \sum_{\langle x, \lambda_x \rangle \in S \, : \, \lambda_x = \lambda} \omega(x \,|\, x_0, S). \tag{2.16}$$

Related to this approach are extensions of NN classification which make use of fuzzy sets [27, 33, 216, 218]. By weighting neighbors according to their distance, these methods compute a "fuzzy" classification

$$\lambda_{x_0}^{est} = \big( u_{\lambda_1}(x_0), \ldots, u_{\lambda_m}(x_0) \big) \tag{2.17}$$

for a new instance $x_0$. That is, $x_0$ is not assigned a unique label in an unequivocal way. Rather, a degree of membership, $u_\lambda(x_0)$, is specified for each label $\lambda$. Consider as an example the fuzzy $k$NN algorithm proposed in [218]. The degree to which $x_0$ is assigned the label $\lambda_\imath$ (is classified into the $\imath$-th class) is given by

$$u_{\lambda_\imath}(x_0) = \frac{\sum_{\jmath=1}^{k} u_{\imath\jmath} \, |x_0 - x_\jmath|^{-2/(m-1)}}{\sum_{\jmath=1}^{k} |x_0 - x_\jmath|^{-2/(m-1)}}, \tag{2.18}$$

where $u_{\imath\jmath} = u_{\lambda_\imath}(x_\jmath)$ is the membership degree of the instance $x_\jmath$ in the $\imath$-th class. The possibility of assigning fuzzy membership degrees $u_{\imath\jmath}$ to labeled instances $x_\jmath$ is seen as a decisive feature. Turning the (non-fuzzy) label $\lambda_{x_\jmath}$ of an observed instance $x_\jmath$ into a fuzzy label allows one to adjust the influence of that instance if it is not considerded prototypical of its class. The constant $m$ in (2.18) determines the weighting of the distance between $x_0$ and its neighbors.

Clearly, (2.18) still has a probabilistic flavor since degrees of membership add up to 1.[23] However, the use of fuzzy labels makes it more general than (2.16). In fact, a fuzzy classification (2.17) can be written as

$$u_{\lambda_0}(x_0) \propto \sum_{\imath=1}^{n} u_{\lambda_0}(x_\imath) \cdot \omega(x_\imath \,|\, x_0, S).$$

Formally, the main difference between a probabilistic estimation and a fuzzy classification is hence the use of fuzzy labels in the latter approach: In the probabilistic case, an observed instance $\langle x, \lambda_x \rangle$ supports the label $\lambda_x$ only. Depending on the "typicality" of the instance (it might concern a "boundary case" whose labeling was not unequivocal), it may also support labels $\lambda \neq \lambda_x$ in the case of fuzzy classification. We shall return to fuzzy-set based approaches of that kind in Chapter 5.

[22] Note, however, that (2.15) actually considers more than $k$ instances if the $k$-th nearest neighbor is not unique. See [288] for an alternative type of distance-weighting in $k$NN which unifies classification and density estimation.

[23] Formally, (2.18) might hence be interpreted as a probability distribution as well. It should be noted, however, that this interpretation might be criticized since the derivation of (2.18) does not assume an underlying probabilistic model.

### 2.2.2 Instance-based learning

Instance-based learning (IBL) algorithms, which belong to the class of *lazy* (supervised) machine learning methods (cf. Section 2.1), are incremental variants of the NN algorithm.[24] They are inspired by exemplar-based models of categorization which have been developed in cognitive psychology [351]. IBL classifies instances based on the assumption that "similar instances have similar classifications." The simplest IBL algorithm, known as IB1 [11], mainly differs from the NN algorithm in that it normalizes the (numeric) attribute values of instances (which are characterized by means of an attribute–value representation), processes instances incrementally, and uses a simple method for tolerating missing attribute values. IB2 extends IB1 by using an editing strategy, i.e., it maintains a memory (case base) of selected cases called prototypes (falsely classified points are added as references). A further extension, IB3, aims at reducing the influence of noisy observations.[25] To this end, a classification record is maintained, which counts the correct and incorrect votes of the stored references. By weighting attribute values in the computation of the distance measure, IB4 and IB5 [5] take the relevance of features into account. The weights are adapted each time a new classification has been made.

Not only have IBL algorithms been used for estimating (discrete) class labels (i.e., for classification), they have also been employed for predicting real-valued attributes (i.e., for regression and function approximation) [221, 420]. Further improvements of IBL algorithms include the incorporation of tolerance toward noisy instances [10], the elimination of irrelevant features [7, 224, 345], the weighting of features [396] or instances [327], approaches to dealing with novel attributes [5], and the consideration of (class-dependent) misclassification costs [290, 364].

IBL algorithms basically consist of three components:

- A *similarity function* computes a numeric similarity between instances.
- A *classification function* estimates the class of a newly presented instance, given the similarities between the new instance and the stored examples as well as the classes (and classification performance) of these examples. It yields a complete *concept description* (a mapping which assigns classes to instances) when being applied to all (still unclassified) instances.
- After each classification task, a *concept description updater* derives a modified concept description by maintaining the memory of instances. The decision whether to retain or remove an instance is based on records of previous classification performance and the information provided by the new classification task.

[24] Though the idea of incremental learning is also contained in basic NN algorithms.

[25] See also [397] for an early work along these lines.

As for the basic NN rule, some efforts have been made to improve the performance of IBL algorithms. Important points, some of which have already been mentioned above, include conceptual aspects such as the reduction of storage requirements by editing and prototype selection [264], the toleration of noise [11], the definition of similarity functions [399], and feature weighting or selection [396], as well as practical issues such as efficient techniques for indexing training examples [394].

### 2.2.3 Case-based reasoning

Case-based reasoning (in a narrow sense, cf. Chapter 1) is one of the more recent developments[26] in AI research and has now become an important and widely applied problem solving technology [234, 315]. It is based on the assumption that "similar problems have similar solutions," another version of the CBI hypothesis (cf. Chapter 1). In fact, this assumption is the guiding principle underlying most CBR systems. More precisely, the idea of CBR is to exploit the experience gained from similar problems in the past and to adapt then successful solutions to the current situation. In order to realize this idea, a CBR system has to maintain (at least) a structured memory of cases (also called a case base) which represents the experience and a means for specifying the similarity between cases (cf. Section 2.3). The basic notion of a *case* is thought of as a representation of knowledge about a specific situation or episode (an episodic chunk of knowledge). In its standard form it consists of two parts, namely a *problem description* and an associated *solution*. The concepts of *problem* and *solution* are very general in nature and have no universally valid definition. Rather, their meaning depends on the respective application.

Case-based reasoning is strongly related to the field of *cognitive modeling*.[27] Indeed, CBR has its origin in the cognitive model of *scripts* [333] and *dynamic memory models* of cognition [332], and it has always been motivated by the idea of providing computational models which are closer to psychology than traditional AI methods.

At a formal level, CBR is built upon the principles of IBL, but involves more complex data structures. In fact, CBR can be seen as an instance-based approach in which instances are complex objects (= cases) rather than points in a Euclidean space and which goes beyond classification as a problem solving task. In comparison to IBL, this additional complexity makes inference more difficult and necessitates further system features, such as the efficient organization of the case base, case retrieval techniques and methods of case adaptation. Broadly speak-

[26] The first and by now well-known CBR conference (DAPRA [233]) was held in the USA in 1988. The first European conference was held in 1993 [314], and the first international conference took place in 1995 in Lisbon, Portugal [382].

[27] The relation between CBR and cognitive science is strongly developed in the USA, whereas in Europe CBR is typically seen as a more technical discipline related to computer science and AI.

ing, inference in IBL is only based on observed cases and a similarity relation,[28] whereas CBR systems also incorporate general domain knowledge.

Case-based reasoning research has largely focused on issues such as the organization of case bases, the efficient retrieval of cases, the assessment of the similarity of cases, and the adaptation of past solutions to the current problem. Considerable research efforts have also been motivated by real-world problems and, hence, have been relatively application-oriented. Until recently, however, only few attempts have been made at formalizing the process of similarity-based inference and its underlying assumptions in a systematic way [101, 99, 141, 199, 296, 105, 201].

**The structure of CBR systems.** A widely accepted CBR methodology which is realized by many practical systems is characterized by the so-called "CBR cycle." It reflects the main components necessary for realizing case-based reasoning, namely the maintenance, the retrieval, the intelligent use, and the update of experiences. The (informal) $R^4$ model of the CBR cycle consists of the following phases [1]:

– RETRIEVE the case(s) from the memory which is (are) most similar to the target problem,

– REUSE the information provided by this case in order to solve the new problem,

– REVISE the proposed solution according to the special requirements of the new problem,

– RETAIN the new experience obtained in the current problem solving episode for future problem solving.

The above outline gives an idea of the principal factors which determine the efficiency of a CBR system, such as

– methods for maintaining and organizing the case base (a simple but common structure is that of a *flat case base* which involves the comparison of the new problem with each case in the memory; more advanced approaches make use of *hierarchical* structures),

– case indexing and case retrieval techniques (indexing means assigning indices to cases for future retrieval and comparison),

– the formalization of the similarity concept (generally in the form of numerical measures, see Section 2.3),

– methods of case adaptation (i.e., the adaptation of solutions of retrieved cases to the problem at hand).

Besides, different types of learning can be incorporated into a CBR system. From a very general perspective, every change of the system in response to its environment can be considered as learning. Thus, the simplest type of learning is

[28] Such methods are called "casuistic" in [296].

perhaps that of storing a new case. Other aspects of learning include the acquisition or adaptation of similarity measures, the learning of retrieval knowledge or the improvement of case adaptation [12].

In its basic form, CBR puts the above model into action by simply retrieving the most similar case from the memory and by using this case for solving the new problem. Since the CBI hypothesis is thus realized in a more or less implicit way (and only for selecting a case from the memory), it is perhaps not astonishing that an explicit formalization has hardly received an attention as yet.

**Applications.** The performance task which is perhaps most often considered in CBR is that of *classification*. Based on the assumption that a case consists of a problem description in the form of a set of symptoms or features, and a class to which it belongs, the task is to determine the class of a new problem. This type of CBR is in fact very close to IBL. However, CBR has also been applied to a wide range of other tasks for which case knowledge plays an important role, such as configuration, diagnosis, decision support, design, and planning [249].

The above types of application (as well as the representation of a case in the form of a *problem* and a *solution*) clearly stress the aspect of problem solving. Still, let us mention that CBR can also be used for other purposes. In *interpretative* CBR, for instance, the focus is on arguing whether a new situation should be treated in the same way as a previous one [80]. Again, the similarity between the two situations plays a crucial role. Needless to say, a clear differentiation between problem solving CBR and interpretative CBR is often not possible in practice. In fact, most CBR systems combine aspects of both types.

**Integration with other techniques.** Hybrid representations combining different paradigms belong to the relatively recent developments in intelligent systems research. In this connection, it is interesting to mention that CBR (or principles thereof) can well be integrated with other methods. Particularly, the combination of CBR and methods of rule induction has led to several interesting approaches [89, 90, 174, 29]. Rule-based reasoning has been used, e.g., for supporting the adaptation task in CBR [246], for assessing similarity measures [14], and for guiding the search and matching process in the retrieval task. CBR can also play the role of a supporting technique in rule induction [346]. Besides, more balanced approaches have been developed in which CBR and rule-based techniques support each other in a common learning and problem solving environment [61]. An interesting architecture in which cases are used for handling exceptions to approximately correct rules has been proposed in [175]. A combined approach is particularly advocated by the complementary properties of the two techniques, namely the representation of general knowledge of a domain in rule induction and the representation of domain-specific knowledge in the form of observed cases in CBR. A thorough elaboration of the potential to integrate CBR with soft computing techniques has more recently been given in [285]. Another current research topic is the use of CBR in reinforcement learning [159].

## 2.3 The concept of similarity

The problem of measuring a kind of similarity or, alternatively, determining a distance between (pairs of) objects naturally arises in many fields of theoretical and application-oriented research. Correspondingly, a large number of specialized similarity or distance measures can be found in literature, such as distance (similarity) measures for vectors, sets, probability measures, sequences, or graph-structured objects. This section is meant as a brief introduction to the concept of similarity from the viewpoint of case-based reasoning and fuzzy set theory, two fields in which similarity plays an important role. We shall not go into too much detail, however, and refrain from a systematic discussion of particular measures, as this is not necessary for subsequent chapters.

### 2.3.1 Similarity in case-based reasoning

As suggested by the CBR cycle, the retrieval of cases which are similar to the new query is an important step of the overall problem solving process. This process hence assumes a quantification of the similarity between objects. In fact, the efficiency of case-based problem solving crucially depends on the adequacy of this quantification. Ideally, the similarity-guided retrieval process provides a case which is *useful* in the sense that the associated solution can easily be adapted to the new problem.

Let $X$ be an arbitrary class of objects. Typically, the similarity between two objects $x, y \in X$ is expressed in terms of a (non-negative) real number $\sigma(x, y)$, i.e., similarity is formalized as a real-valued function $\sigma : X \times X \longrightarrow \mathfrak{R}_{\geq 0}$.[29] The latter is also called a *similarity measure* or *similarity function*. Yet, similarity can also be formalized by means of a relational approach. Indeed, a relation $R \subseteq X^4$ with the intended meaning that

$$(x, y, u, v) \in R \;\Leftrightarrow\; x \text{ is at least as similar to } y \text{ as } u \text{ to } v$$

already allows for defining a *nearest neighbor* of an object $x$:

$$\mathsf{NN}(x, z) \overset{\text{df}}{\Leftrightarrow} \forall\, y \in X \;:\; (x, z, x, y) \in R.$$

The relational approach hence suffices for realizing the CBR process outlined above. This also shows that CBR does generally not assume a cardinal interpretation of a similarity measure. In fact, what is important is only the *order relation* between degrees of similarity.

From a mathematical point of view, *similarity* and *distance* can be seen as dual concepts [371, 378]. A similarity function $\sigma$ and a distance measure $\Delta$ (also defined over $X$) are *compatible* if $R_\sigma = R_\Delta$, where the relation $R_\sigma$ is induced by $\sigma$ via

[29] One could also think of relaxing the assumption that $\sigma \geq 0$, but such measures are usually not considered.

$$(x, y, u, v) \in R_\sigma \overset{\text{df}}{\Leftrightarrow} \sigma(x, y) \geq \sigma(u, v),$$

and $R_\Delta$ is defined in an analogous way.

There are a number of (more or less reasonable) properties which may be required of a similarity function $\sigma$, notably reflexivity, symmetry, and transitivity (the latter being expressed in terms of a related distance $\Delta$):

- $\forall x \in X \, : \, \sigma(x, x) = 1$,
- $\forall x, y \in X \, : \, \sigma(x, y) = \sigma(y, x)$,
- $\forall x, y, z \in X \, : \, \Delta(x, z) \leq \Delta(x, y) + \Delta(y, z)$.

All these properties are discussed controversially in literature, however. Several authors argue in favor of reflexivity and symmetry, whereas TVERSKY claims that both properties are too strong [373]. His main argument against symmetry relies on a differentiation between a *subject* and a *referent*. For instance, people generally find an ellipse (the subject) more similar to a circle (the referent) than vice versa. On the one hand, TVERSKY's argument is quite convincing. Besides, it is confirmed by a number of examples and experimental studies. On the other hand, however, it shows that a careful distinction between similarity as a *descriptive* and similarity as a *normative* concept has to be made. From a descriptive point of view, which is clearly dominant in cognitive psychology, it seems that similarity should indeed be considered as an asymmetric concept. From a normative point of view, however, one might well argue that symmetry is a reasonable requirement. Similar arguments also apply in connection with the property of reflexivity.

The objections which can be raised to transitivity are more convincing, even from a normative point of view. For instance, the $\{0, 1\}$-valued measure which assigns a similarity of 1 to real numbers $x, y$ iff $|x - y| < \varepsilon$ might be advocated as being useful for certain applications, but it is obviously not transitive. (As will be seen below, weaker concepts of transitivity can be handled within the framework of fuzzy sets.)

An interesting framework in which different classes of similarity measures are distinguished has been proposed in [47]. Each of the classes is identified by some requirements a measure must satisfy, and specific applications it might be useful for: Measures of *similitude* are non-symmetrical and quantify the extent to which an object comes close to a reference object. Measures of *inclusion*, also non-symmetrical, quantify the degree to which an object can be considered as a special case of a reference object. Measures of *resemblance* do not assign specific roles to the objects under consideration and are hence symmetric.

The computation of similarity largely depends on the representation of objects. Several approaches have been developed, including

- the *feature-based approach*, where an object is represented by a set of features (properties that apply to the object) and similarity is derived from the commonality or difference of the features associated with two objects [373],
- the *geometric approach*, in which objects are coded as points in some $n$-dimensional (metric) space and similarity is inversely related to distance [341],
- the *structural approach*, in which the relation between cases is represented by means of a graph structure and similarity is based on graph matching [58].

Anyway, the most common approach is to make use of an attribute–value representation, i.e., to characterize a case (input, output) as a vector $a = (a_1, \ldots, a_n)$ of attribute values. Denote by $\mathcal{A}_\imath$ the domain of the $\imath$-th attribute and let $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_n$. Moreover, suppose a *global similarity measure* $\sigma : \mathcal{A} \times \mathcal{A} \longrightarrow \mathfrak{R}_{\geq 0}$ to be given. The so-called *local-global principle* makes the following assumption: There are *local similarity measures* $\sigma_\imath : \mathcal{A}_\imath \times \mathcal{A}_\imath \longrightarrow \mathfrak{R}_{\geq 0}$ and a *composition function* $f : (\mathfrak{R}_{\geq 0})^n \longrightarrow \mathfrak{R}_{\geq 0}$ such that

$$\sigma(a, a') = f\left(\sigma_1(a_1, a'_1), \ldots, \sigma_n(a_n, a'_n)\right) \tag{2.19}$$

for all $a = (a_1, \ldots, a_n), a' = (a'_1, \ldots, a'_n) \in \mathcal{A}$.[30] There are some reasonable properties which might be assumed in connection with the representation (2.19). For example, the *global monotonicity axiom* states that

$$\sigma(a, a') < \sigma(a, a'') \Rightarrow \exists \imath \in \{1, \ldots, n\} \, : \, \sigma_\imath(a_\imath, a'_\imath) < \sigma_\imath(a_\imath, a''_\imath).$$

for all $a, a', a'' \in \mathcal{A}$.

EXAMPLE 2.2. Commonly used similarity measures are often derived from the weighted Euclidean metric

$$\mathit{\Delta} : (a, a') \mapsto \left( \sum_{\imath=1}^{n} \omega_\imath \cdot (a_\imath - a'_\imath)^2 \right)^{1/2} .$$

This measure obviously assumes numeric attributes. An example of a more general similarity measure is

$$\sigma : (a, a') \mapsto 1 - \sum_{\imath=1}^{n} \omega_\imath \cdot \mathit{\Delta}_\imath(a_\imath, a'_\imath),$$

where $\mathit{\Delta}_\imath$ is a *normalized* Euclidean distance in the case of numeric attributes and the discrete distance measure (which takes the value 0 if $a_\imath = a'_\imath$ and 0 otherwise) in the case of categorical variables (the $\omega_\imath$ are non-negative weights such that $\omega_1 + \ldots + \omega_n = 1$). □

[30] Similar assumptions on the decomposition of a high-dimensional measure into several low-dimensional ones are made in utility theory.

It should be mentioned that more general formalizations of similarity can be obtained by weakening the assumption that a measure is numeric [213]. Moreover, it is not compulsory to measure similarity on a completely ordered scale, as already suggested by the relational approach. For instance, a similarity measure might be a $X \times X \longrightarrow \mathcal{L}$ mapping with $\mathcal{L}$ being a lattice structure [283, 56]. See [252, 399] for an overview and a comprehensive analysis of special similarity (distance) measures used in CBR.

It goes without saying that (case-based) reasoning and problem solving will often turn out as inadequate from a reuse perspective if similarity is simply derived as a function of certain properties of objects, e.g., features or attribute values. In other words, it might not be appropriate to assume that the case which is most similar in the sense of a "surface measure" of this kind can easily be adapted to fit the target problem. Consequently, similarity measures should be augmented by deeper, domain-specific knowledge about the adaptability of cases, at least if CBR is realized in its basic form (namely by retrieving the most similar case and by adapting the corresponding solution to the new problem). The adaptation-guided retrieval technique proposed in [354], for instance, improves the efficiency and accuracy of case-retrieval by means of an algorithmic measure of adaptability.

The CBI hypothesis is trivial (circular) when saying that two problems are similar if the related solutions are similar, a fact which is often criticized in the philosophical and psychological literature [178]. As already pointed out in Chapter 1, however, this objection to CBR as a useful reasoning principle is hardly relevant from a practical point of view: Since the solution to the target problem is unknown, this "ideal" similarity measure cannot be computed. Rather, the objective is to define similarity between problems (without knowing the solutions) in such a way that the CBI hypothesis holds at least approximately. The more knowledge about the domain is available, the better a similarity measure will generally meet this requirement [160]. The problem of *similarity assessment*, i.e., the learning and adaptation of similarity measures, is one of the central topics in CBR research [358]. Learning is usually realized by adapting the parameters of a (parameterized) similarity function. So-called *feature weighting* methods [395, 396] can be mentioned as a typical example.

In this connection, let us make a remark on the semantics of the similarity concept. In CBR, it is sometimes proposed to interpret similarity in terms of other concepts, such as (gradual) truth or probability [312]. For instance, the idea to define the similarity between two problems as the *probability* that the associated solutions are similar (identical) can be seen as a straightforward generalization of the above-mentioned "ideal" similarity measure. However, we prefer to consider similarity as a (cognitive) concept in its own right. In fact, it can hardly be denied that similarity plays an independent role in human reasoning, as do related concepts such as belief and preference. In this sense, the cognitive basis of similarity-based reasoning is clearly undermined when reducing similarity to other concepts. Besides, it should be noted that a similarity measure might be ideal in the above-mentioned sense but, at the same time, rather counter-intuitive. In this respect, one might think of expressing the CBI hypothesis in a more restrictive way as follows: “There are *intuitively reasonable* measures of similarity $\sigma, \sigma'$ such that two inputs are similar in the sense of $\sigma'$ if the related outputs are similar in the sense of $\sigma$.” This hypothesis is indeed non-trivial, since measures which are ideal and reasonable at the same time need not necessarily exist.

### 2.3.2 Similarity and fuzzy sets

The concept of similarity is also closely related to the theory of fuzzy sets [412, 284]. In fact, one of the main semantics of the membership function of a fuzzy set $A \subset X$ is that of encoding degrees of similarity between elements of $X$ and elements which are prototypical (and for which the degree of membership in $A$ is 1) of a certain (fuzzy) concept characterized by $A$ [26, 320]. In the same way as the idea of a fuzzy subset generalizes that of a classical set, the concept of *similarity* can be seen as a generalization of the classical notion of *equivalence*. That is, a similarity relation can be interpreted as a generalization of an equivalence relation. By taking fuzzy equivalence, also referred to as indistinguishability, as a basic concept, it is even possible to view fuzzy sets as an induced concept [225]. This point of view suggests a way to provide meaningful semantics for certain fuzzy reasoning schemes. In [226], for instance, fuzzy control has been interpreted as interpolation in the presence of indistinguishability. Likewise, “fuzzy granules” are considered as “objects forming a granule drawn together by similarity” in connection with the modeling of fuzzy graphs [31].

Let $\top$ be a triangular norm (t-norm), i.e., a function $\top : [0,1] \times [0,1] \longrightarrow [0,1]$ which is associative, commutative, nondecreasing in both arguments, and such that $\top(x,1) = x$ for all $0 \leq x \leq 1$ [227]. A $\top$-similarity on a set $X$ is a fuzzy relation $\sigma : X \times X \longrightarrow [0,1]$ satisfying reflexivity, symmetry, and $\top$-transitivity:

- $\forall\, x \in X \,:\, \sigma(x,x) = 1$,
- $\forall\, x, y \in X \,:\, \sigma(x,y) = \sigma(y,x)$,
- $\forall\, x, y, z \in X \,:\, \top(\sigma(x,y), \sigma(y,z)) \leq \sigma(x,z)$.

Depending on the choice of $\top$, the assumption of transitivity turns out to be more or less restrictive. In fact, it is maximally restrictive for the largest t-norm, namely the minimum operator. It is minimally restrictive for the drastic product, which is given by the mapping

$$\top : (x,y) \mapsto \begin{cases} \min(x,y) & \text{if} \quad \max(x,y) = 1 \\ 0 & \text{if} \quad \max(x,y) < 1 \end{cases} .$$

A similarity measure $\sigma$ is called *separating* if $\sigma(x,y) = 1 \Leftrightarrow x = y$ holds true for all $x, y \in X$. (The same property is denoted *strong reflexivity* in case-based reasoning.) A relation which satisfies reflexivity and symmetry is called a *proximity relation* [100].

Suppose a similarity relation $\sigma : X \times X \longrightarrow [0,1]$ to be given. Then, each element $u \in X$ induces a *fuzzy equivalence class* $[u]_\sigma$, namely the fuzzy set of elements close to $u$. This fuzzy subset of $X$ is characterized by the membership function $x \mapsto \sigma(x,u)$.

As already mentioned above, there is a close relation between the notions of similarity and distance. A mapping $\Delta : X \times X \longrightarrow [0, \infty]$ is an extended (real-valued) pseudometric on $X$ if

- $\forall x \in X \, : \, \Delta(x,x) = 0$,
- $\forall x,y \in X \, : \, \Delta(x,y) = \Delta(y,x)$,
- $\forall x,y,z \in X \, : \, \Delta(x,y) \leq \Delta(x,z) + \Delta(z,y)$,

where $x + \infty = \infty + x = \infty$ for all $x \in [0,\infty]$ by definition. Let $f : [0,1] \longrightarrow [0,\infty]$ be a continuous and strictly decreasing function satisfying $f(1) = 0$. The function

$$\sigma : X \times X \longrightarrow [0,1] \, , \, (x,y) \mapsto f^{(-1)}(\Delta(x,y))$$

is then a $\top$-similarity on $X$, where $f^{(-1)}$ denotes the pseudoinverse

$$f^{(-1)} : [0,\infty] \longrightarrow [0,1] \, , \, x \mapsto \begin{cases} f^{-1}(x) & \text{if} \quad x \in f([0,1]) \\ 0 & \text{if} \quad x \notin f([0,1]) \end{cases} ,$$

and $\top$ is the continuous Archimedean t-norm[31] generated by $f(\cdot)$. In this case, we shall say that $\sigma$ is *$\Delta$-related* (via $f(\cdot)$). The other way round, the function $\Delta : X \times X \longrightarrow [0,\infty]$ defined by $\Delta(x,y) = f(\sigma(x,y))$ is an extended pseudometric if $f(\cdot)$ is an additive generator of a continuous Archimedean t-norm $\top$, i.e., if $\top$ is given by the mapping

$$(x,y) \mapsto f^{(-1)}(f(x) + f(y)),$$

and if $\sigma$ is a $\top$-similarity on $X$.

### 2.3.3 Aggregation of local similarity measures

According to the above-mentioned local-global principle (page 35) for similarities, deriving a global similarity relation from a set of individual relations comes down to defining an adequate aggregation operator. Ideally, such an operator should preserve certain properties of the individual relations. Most aggregation operators do preserve reflexivity and symmetry, but not necessarily transitivity. This remark does already apply to the simple arithmetic mean, i.e., the measure $\sigma$ with $f(\cdot)$ in (2.19) given by

$$f : (x_1, \ldots, x_n) \mapsto \frac{1}{n} \sum_{\imath=1}^{n} x_\imath . \tag{2.20}$$

[31] A t-norm $\top$ is called Archimedean if $\top(x,x) < x$ for all $0 < x < 1$.

Still, for a t-norm $\top$ it is not difficult to show that the global measure

$$\sigma : (a, a') \mapsto \sigma_1(a_1, a_1') \top \sigma_2(a_2, a_2') \top \ldots \top \sigma_n(a_n, a_n')$$

is $\top$-transitive whenever the same holds true for the local measures $\sigma_\imath$ $(1 \leq \imath \leq n)$.

Simple examples of aggregation operators which have been studied extensively in the literature on fuzzy sets are the minimum and the maximum. *Weighted* aggregations [151] are of particular interest in connection with CBI since they allow for assigning a level of importance to the attributes.

In this connection, it should be noted that literature offers weighted aggregation operators other than the weighted arithmetic mean. Such operators are important when dealing with non-numeric attributes or ordinal scales of similarity (cf. Chapter 5) where averaging does not make sense [112]. As an example consider the following similarity function which is a weighted version of the min-operator:

$$\sigma : (a, a') \mapsto \min_{1 \leq \imath \leq n} \max\{\sigma_\imath(a_\imath, a_\imath'), 1 - \lambda_\imath\} \tag{2.21}$$

with $0 \leq \lambda_\imath \leq 1$ and $\max_{1 \leq \imath \leq n} \lambda_\imath = 1$. The value $\lambda_\imath$ defines the level of importance of the $\imath$-th attribute. As can be seen, $\lambda_\imath = 1$ corresponds to full importance, whereas $\lambda_\imath = 0$ means that the $\imath$-th attribute is completely ignored. Let us mention that (2.21) preserves reflexivity, symmetry, and transitivity.

So-called *ordered weighted averaging* (OWA) operators, which generalize several well-known operators, such as the average and the minimum, have also been proposed as aggregation operators in the context of case-based reasoning [409]. Another interesting aggregation operator (apparently not yet considered in connection with similarity evaluation) is the Choquet integral [67, 180]. Not only does this operator allow for weighting individual attributes, it can also take interdependencies between attributes into account. The Choquet integral of an extended real-valued function $f(\cdot)$ on a topological space $\Omega$ is defined as

$$\int^{ch} f \, d\eta \overset{\text{df}}{=} \int_0^\infty \eta([f > t]) \, dt + \int_{-\infty}^0 (\eta([f > t]) - 1) \, dt,$$

where $[f > t] = \{\omega \in \Omega \,|\, f(\omega) > t\}$ and $\eta$ is a so-called *capacity* (a special type of set function).

For our purpose, it suffices to consider the Choquet integral for the finite case, namely for $\Omega = \{1, \ldots, n\}$. Let $\sigma_1, \ldots, \sigma_n$ be local similarity measures. By making use of the Choquet integral as an aggregation operator we obtain the following global measure:

$$\sigma : (a, a') \mapsto \int^{ch} h \, d\eta, \tag{2.22}$$

where $h : \{1, \ldots, n\} \longrightarrow [0, 1]$ is given by the mapping $\imath \mapsto \sigma_\imath(a_\imath, a_\imath')$. Thus, $h(\imath)$ denotes the similarity between the attribute values $a_\imath$ and $a_\imath'$. Moreover,

$\eta : 2^{\Omega} \longrightarrow [0,1]$ is a normalized and inclusion-monotone measure, i.e., $\eta(\emptyset) = 0$, $\eta(\Omega) = 1$, and $\eta(A) \leq \eta(B)$ for $A \subset B \subset \Omega$.

Let $\pi$ denote a permutation of $\{1 \ldots n\}$ such that $h(\pi(\imath)) \leq h(\pi(\imath + 1))$ for $1 \leq \imath < n$. That is, $\pi$ arranges the attributes according to the degree of similarity. The similarity function (2.22) can then be written as follows:

$$\sigma : (x, x') \mapsto \sum_{\imath=1}^{n} h(\pi(\imath)) \cdot \left(\eta(\{\pi(1) \ldots \pi(\imath)\}) - \eta(\{\pi(1) \ldots \pi(\imath - 1)\})\right), \tag{2.23}$$

where $\eta(\emptyset) \overset{\mathrm{df}}{=} 0$. Note that (2.23) includes several known aggregation operators as special cases. For instance, with $\eta$ being the counting measure $X \mapsto 1/n \cdot |X|$ we obtain the arithmetic mean. More generally, let $\eta$ be the additive measure with $\eta(\{\imath\}) = \alpha_\imath$ for all $1 \leq \imath \leq n$, where $0 \leq \alpha_\imath \leq 1$ and $\alpha_1 + \ldots + \alpha_n = 1$. The global measure (2.23) is then given by the weighted arithmetic mean

$$\sigma : (x, x') \mapsto \sum_{\imath=1}^{n} \alpha_\imath \cdot \sigma_\imath(a_\imath, a'_\imath).$$

OWA operators are recovered if $\eta(\cdot)$ is symmetric (commutative), i.e., if $\eta(X)$ only depends on the cardinality of $X$. For example, a kind of threshold similarity can be modeled by letting, for a fixed $k \in \{1 \ldots n\}$, $\eta(X) = 1$ if $|X| \geq n - k + 1$ and $\eta(X) = 0$ otherwise. The similarity between two objects is then given by the $k$-th highest among the similarity degrees, that is,

$$\sigma(a, a') = h(\pi(n - k + 1)), \tag{2.24}$$

expressing that the objects must resemble each other according to "at least $k$ out of $n$" criteria. The special case $k = n$ yields the minimum operator as an aggregation function:

$$\sigma : (a, a') \mapsto \min_{1 \leq \imath \leq n} \sigma_\imath(a_\imath, a'_\imath). \tag{2.25}$$

As already mentioned above, an interesting aspect in connection with the Choquet functional as an aggregation operator is its capability to take interdependecies between different attributes into account. In fact, in many applications the global similarity between two objects does not simply correspond to the (weighted) sum of the local similarities. Suppose, for example, that the $\imath$-th and the $\jmath$-th attribute are complementary in a certain sense. In order to call two objects similar, it might hence be required that *both*, $a_\imath$ is similar to $a'_\imath$ and $a_\jmath$ is similar to $a'_\jmath$. The minimum in (2.25), for instance, might be seen as an adequate aggregation operator if *all* attributes are complementary in this sense. The measure (2.24) combines this type of complementarity of attributes with a compensation effect, since the similarity with regard to one attribute can compensate for the dissimilarity with respect to another one.

The Choquet integral can be seen as a generalized (weighted) mean value operator.[32] If similarity is measured on an ordinal scale, the Choquet integral can be replaced by the Sugeno integral [362]. The latter defines a generalization of an alternative location parameter, namely the median which can also be used within a qualitative setting.

Again, let $\eta$ be a normalized and monotone measure and denote by $f(\cdot)$ a measurable $\Omega \longrightarrow \mathfrak{R}_{\geq 0}$ function. The Sugeno integral of $f(\cdot)$ with respect to $\eta$ is defined as follows:

$$\int^{su} f \, d\eta \overset{\mathrm{df}}{=} \sup_{0 \leq \alpha \leq 1} \min\{\alpha, \eta([f > \alpha])\}. \tag{2.26}$$

As can be seen, Sugeno's integral is formally obtained by replacing addition and multiplication in the classical Lebesgue integral by the supremum and infimum, respectively. Applying (2.26) within our context yields the global similarity measure

$$\sigma : (a, a') \mapsto \max_{1 \leq \imath \leq n} \min \left\{ h_{\pi(\imath)}, \eta(\{\pi(1), \ldots, \pi(\imath)\}) \right\}. \tag{2.27}$$

This section has only given a first idea of how to make use of generalized measures and integrals in the context of similarity evaluation. Of course, there are questions of practical importance which call for further investigation. In particular, this concerns the definition (elicitation) of the measure $\eta$ [179]. How should an expert determine $\eta$ to depict his view of similarity in an optimal way? Besides, it would be interesting to solve the *inverse problem*: Given a set of examples in the form of global similarity evaluations provided by some expert, induce (or approximate) the measure $\eta$ this expert has used in order to derive these evaluations. Likewise, given a set of training examples, one might try to adapt the measure $\eta$ so as to maximize the performance of a CBI method, e.g., the predictive accuracy of an NN classifier.

## 2.4 Case-based inference

In subsequent chapters, we shall propose several models of similarity-based (case-based) inference which are based on explicit formalizations of the CBI hypothesis. As already said, we concentrate on *prediction* as a performance task [99, 100, 142, 265],[33] which is in line with the idea underlying case-based learning algorithms [4], exemplar-based reasoning [220, 327], memory-based reasoning [359], and instance-based learning [11].[34] Thus, we consider the task of exploiting

[32] It is used as such in non-additive expected utility.

[33] Case-based prediction has already been applied to different domains, e.g., to real estate property appraisal [176, 40] and the forecasting of power load [208] and retail sales [263].

[34] These methods can be seen as non-generalization approaches to the concept learning problem addressed in machine learning (cf. Sections 2.1 and 2.2). They are generally concerned with a classification task, i.e., the prediction of the *class* to which a case belongs.

past experience represented by a memory of previously observed cases – against background knowledge in the form of the CBI hypothesis – in order to predict or characterize the output of a new (query) input. This is what we shall subsequently understand by *case-based inference* (CBI).

In connection with case-based reasoning, CBI essentially concerns the RETRIEVE (and REUSE) processes within the $R^4$ model of the CBR cycle [1]. In fact, CBI does not cover the complete process of (case-based) problem solving, i.e., it will generally not return the ultimate solution to a new problem. Rather, it is intended to bring a promising set of solutions into focus. This way, CBI supports subsequent stages of the overall problem solving process, in the sense that these stages can then focus on the most promising candidates. These stages, which roughly correspond to the REVISE part of the $R^4$ model, are often not directly "case-based" but make use of domain-specific knowledge, user input, or further problem solving strategies. This can be exemplified by the combination of *case-based* and *generative* planning: The plans which have been retrieved from the memory and suggested by a case-based planner for solving a new problem are used as a source of modification by a generative planning method [381]. A further example is that of using similarity-based predictions in order to restrict search spaces in heuristic search (cf. Chapter 8).

According to the point of view adopted above, case-based inference has important aspects in common with statistical (prediction) methods and, more generally, with approaches to machine learning. Namely, the main task is defined as one of deriving predictions from observed data. Still, there are at least two aspects in which case-based inference deviates from classical (model-based) approaches, as will become clear in subsequent chapters. Both aspects have already been touched on in Section 2.1. Firstly, CBI delays the processing of training examples until a new query instance $x_0$ is received, which qualifies it as a lazy learning method [6]. Secondly, CBI does not form an explicit hypothesis of the target function over the entire *instance space*, as (model-based) eager methods do. Besides, let us also mention that, as a CBR-related inference scheme, CBI is often concerned with the prediction of complex outputs. In contrast, statistical inference and machine learning are more focused on problems such as classification and regression, i.e., the prediction of categorical or numerical outputs.

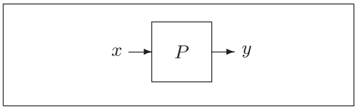


**Fig. 2.2.** A data generating process on a very abstract level.

Nevertheless, concerning the second point Section 2.1 has shown that even instance-based approaches have to incorporate some kind of background knowledge (in the form of a related hypothesis) into the process of generalization. Let us elaborate on this aspect more closely by considering a data generating process, $P$, on a very abstract level, as shown in Fig. 2.2. This process simply transforms an input $x \in X$ into an output $y \in Y$. The ensemble of all such *instances* $(x, y)$ (resp. their joint probability distribution) can be thought of as defining the *system* under consideration. Now, given a certain input $x_0$, the task shall be to predict the associated output $y_0$. To this end, model-based approaches generally make some *structural assumptions* concerning the process $P$. Mathematically, such assumptions are represented by a (parameterized) set $\mathcal{H}$ of functions $h : X \longrightarrow Y$, called the hypothesis space in machine learning [271]. Given a set of data in the form of observed instances, one searches the space $\mathcal{H}$ for the hypothesis $h_0$ which – in a specific sense – fits or reproduces these observations best. A prediction of the output $y_0$ is then given by $\widehat{y}_0 = h_0(x_0)$. As a simple example consider a linear regression model

$$Y = h(X) = \alpha_1 \cdot X_1 + \ldots + \alpha_n \cdot X_n + \varepsilon. \tag{2.28}$$

The main structural assumption about $P$ made by this model is that of a linear relationship between the output $Y$, the inputs $X_1, \ldots, X_n$ (which constitute the input vector $X$) and an error term $\varepsilon$. Thus, the hypothesis space $\mathcal{H}$ consists of all functions of the form (2.28), where $\alpha_k \in \mathfrak{R}$ $(1 \leq k \leq n)$. In general, the structure of $P$ is characterized further by assumptions concerning the statistical distribution of $\varepsilon$. Such assumptions have an essential impact on the definition of criteria for selecting an optimal hypothesis.[35] Assuming a certain statistical distribution for $\varepsilon$, for instance, allows for the selection of a hypothesis according to the Maximum Likelihood principle.

Typically, the hypotheses $h \in \mathcal{H}$ establish a direct relationship between properties (attributes) of the instances $(x, y) \in X \times Y$. The regression model (2.28) illustrates this quite well: The value of a variable, e.g., the (monetary) income of a person, is modeled as a function of several (explaining) properties of that person, such as the age, sex, and education. The CBI hypothesis is obviously of a different kind. In fact, it does not make assumptions about the properties of objects under consideration, but about the *similarity* between such objects. The concept of similarity is *supplementary* in the sense that it is generally not defined a-priori for a certain system. Moreover, it can be seen as a *derived property* which is related to *tuples* of instances. Thus, CBI makes structural assumptions about the process $P$ not directly at the *system* or *instance level* but at the, say, *similarity level*.[36] Seen from this perspective, the process of CBI should mainly take place in some kind of *similarity space* instead of the *instance space*. Consequently, our formalization of CBI will proceed from this level.

[35] Besides, (2.28) would actually be meaningless without such assumptions since $\varepsilon$ could simply be defined (as a function of $x$ and $y$) such that (2.28) holds.

[36] The possibility of utilizing *derived* properties can be of great advantage not only in connection with CBI. Feature generation methods in machine learning, for example, can improve the classification power significantly.

Let us illustrate the different nature of hypotheses concerning the instance level and the similarity level by means of a simple example. To this end, consider some (unknown) function $f : [0,1] \longrightarrow [0,1]$ and suppose the similarity of two numbers $0 \leq x, y \leq 1$ to be defined as $1 - |x - y|$. A hypothesis at the instance level, like a linear relationship, refers to $f(\cdot)$ directly. As opposed to this, a similarity hypothesis concerns the *variation* of $f(\cdot)$, e.g., properties of its derivative. We might assume, for instance, that the similarity of two outcomes, $f(x)$ and $f(x')$, is always greater or equal to the similarity of the respective inputs, $x$ and $x'$. Mathematically, this is nothing else than saying that $f(\cdot)$ is Lipschitz-continuous (with Lipschitz-constant 1).

It should be clear that the expressiveness of a similarity hypothesis strongly depends on the definition of the underlying similarity measures. Consider again some (unknown) function $f : X \longrightarrow Y$ as an example, where $X$ and $Y$ are endowed with a metric $\varDelta$, and let the similarity of inputs, $\sigma$, be related to $\varDelta$. If $\varDelta$ is the discrete metric, the similarity of two inputs $x$ and $x'$ is given by

$$\sigma(x, x') = \begin{cases} 1 & \text{if} \quad x = x' \\ 0 & \text{if} \quad x \neq x' \end{cases} .$$

It will then hardly be possible to express a meaningful hypothesis in terms of similarities. Fortunately, similarity measures will often be much more "discriminating" than in this extreme example (cf. Section 3.3). Exploiting the similarity structure of a system will then lead to an overall gain of information.

Needless to say, the usefulness of different types of (inductive) inference strongly depends on the application at hand. The success of model-based approaches generally requires a relatively simple target function over the instance space.[37] Similarity-based methods might hence be more appropriate if this assumption is not satisfied. Consider a time series like a simple random walk as an example. The function which maps time points $t$ to system states $x(t)$ does not have any simple global structure.[38] Still, it satisfies the CBI assumption in the sense that the distance between states associated with neighbored time points is generally small (cf. Section 3.5).[39] This example also shows that the CBI hypothesis can well be satisfied even though the target function has a rather complex global structure.

Of course, similarity-based inference and model-based reasoning do not exclude each other. That is, exploiting the similarity structure does not prevent us from applying reasoning procedures at the instance level as well. On the contrary, the combination of rule-based reasoning (at the instance level) and case-based

[37] Approximation methods based on neural networks can handle rather complex functions and might hence be seen as an exception. In fact, these (black box) methods do not require the specification of a (simple) hypothesis. Nevertheless, good results will only be obtained with an appropriate network structure. Thus, some background knowledge is still necessary.

[38] Even most model-based approaches do not try to infer the global structure of stochastic processes directly, but rather use models in order to explain the *changes* of system states.

[39] See [276] for the application of CBR in the context of time series prediction.

reasoning, briefly touched on in Section 2.2.3, shows that it can be reasonable to combine corresponding approaches or the respective inference results. In fact, what we shall realize in subsequent chapters can be seen as a combination of similarity-based reasoning and constraint-based, probabilistic or fuzzy set-based inference.

After having outlined the basic ideas underlying CBI in a rather informal way, we are now going to introduce the formal framework from which we shall proceed in subsequent chapters. Within this framework, a distinction between *deterministic* and *non-deterministic* (prediction) problems is made.

### 2.4.1 Deterministic inference problems

In subsequent chapters, we shall adopt parts of the basic CBR terminology. In particular, an *observation*, *sample* or *training example* will often be called a *case* or an *instance*. This is somewhat more general than certain specialized terms such as *pattern* (which is used in pattern recognition and, hence, refers to a concrete application). Nevertheless, apart from the context all these expressions do basically have the same meaning and can be considered as synonyms. A *case* is defined as a tuple consisting of an *input* and an associated *output* or *outcome*, usually denoted by $s$ and $r$, respectively. Again, we prefer these slightly more general expressions to the terms "problem" and "solution" which are commonly used in CBR since we do not focus on problem solving as a performance task.

**Definition 2.3 (deterministic** CBI **setup).** A deterministic CBI setup is defined as a 6-tuple

$$\Sigma = \left\langle\, (\mathcal{S}, \mu_{\mathcal{S}}), \mathcal{R}, \varphi, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, \mathcal{M} \,\right\rangle ,$$

where $\mathcal{S}$ is a countable set of inputs endowed with a probability measure $\mu_{\mathcal{S}}$ (defined on $2^{\mathcal{S}}$), $\mathcal{R}$ is a set of outputs, and $\varphi : \mathcal{S} \longrightarrow \mathcal{R}$ assigns outputs to inputs. The functions $\sigma_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \longrightarrow [0,1]$ and $\sigma_{\mathcal{R}} : \mathcal{R} \times \mathcal{R} \longrightarrow [0,1]$ define similarity measures on the set of inputs and the set of outputs, respectively. $\mathcal{M}$ is a finite memory

$$\mathcal{M} = \big(\langle s_1, r_1\rangle, \langle s_2, r_2\rangle, \ldots, \langle s_n, r_n\rangle\big) \tag{2.29}$$

of cases $c = \langle s, \varphi(s)\rangle \in \mathcal{S} \times \mathcal{R}$.[40] We denote by $\mathcal{M}^{\downarrow}$ the projection of the memory $\mathcal{M}$ to $\mathcal{S}$, i.e., $\mathcal{M}^{\downarrow} = (s_1, \ldots, s_n)$. Moreover,

$$\begin{aligned} D_{\mathcal{S}} &\overset{\text{df}}{=} \big\{\sigma_{\mathcal{S}}(s, s') \,|\, s, s' \in \mathcal{S}\big\} \\ D_{\mathcal{R}} &\overset{\text{df}}{=} \big\{\sigma_{\mathcal{R}}(\varphi(s), \varphi(s')) \,|\, s, s' \in \mathcal{S}\big\} \end{aligned}$$

define, respectively, the set of similarity degrees of inputs and outputs that can actually be attained.[41] □

[40] We shall use the term "CBI setup" also without having defined a fixed memory, in which case it actually refers to the 5-tuple $\langle (\mathcal{S}, \mu_{\mathcal{S}}), \mathcal{R}, \varphi, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}\rangle$.

[41] Note that $\text{card}(\mathcal{S}) \leq \aleph_0$ implies the same to be true for the sets $\mathcal{R}, D_{\mathcal{S}}, D_{\mathcal{R}}$.

A few remarks on Definition 2.3 are in order. The probability measure $\mu_{\mathcal{S}}$ models the occurrence of inputs. Thus, we assume the inputs resp. the associated cases which constitute the memory $\mathcal{M}$ to be chosen repeatedly and independently according to $\mu_{\mathcal{S}}$. This assumption of independent and identically distributed (i.i.d) observations is standard in statistics and machine learning. It is less restrictive than it might appear and should, therefore, not be overrated. Subsequent chapters will show that it is not necessary to make explicit assumptions about $\mu_{\mathcal{S}}$ in connection with similarity-based inference schemes.

In accordance with the above assumptions, we shall treat a memory $\mathcal{M}$ and its projection $\mathcal{M}^{\downarrow}$ not as a set but as a *sequence* of not necessarily different cases (inputs). Still, we retain the standard notations for operations on sets. The meaning of these operations in the new context will generally be obvious. For example, $(s_1, \ldots, s_n) \cup (s'_1, \ldots, s'_m)$ defines the list $(s_1, \ldots, s_n, s'_1, \ldots, s'_m)$ of (not necessarily different) inputs.

We do not make special assumptions on the characterization of inputs or outputs. Utilizing an *attribute–value representation* is common practice in CBR, or AI in general (cf. Section 2.3). That is, inputs as well as outputs are marked as vectors of (not necessarily numeric) attribute values. Yet, other types of representation, for example graphs, are also possible, as long as they allow for a meaningful (and efficient) computation of similarity measures. These measures are assumed to be reflexive, symmetric and normalized in the sense that degrees of similarity are elements of the unit interval $[0, 1]$, where a value of 1 (0) corresponds to perfect (dis)similarity.

The assumption that an input $s \in \mathcal{S}$ determines the associated outcome $r = \varphi(s) \in \mathcal{R}$ (which is the reason for calling a corresponding CBI setup *deterministic*) does not imply that the latter is *known* as soon as the input is characterized. For example, let inputs correspond to instances of a class of combinatorial optimization problems. Moreover, define the output associated with an input as the set of all optimal solutions of the associated problem. Deriving these solutions from the description of the problem might involve a computationally complex process. Moreover, one might think of examples where the mapping $\varphi$ is not even computable. In this connection, similarity-based inference serves as a method which supports the overall process of problem solving by predicting the output associated with a certain input. To this end, CBI performs according to the CBI *principle*: It exploits experience represented by precedent cases, to which it "applies" background knowledge in the form of the heuristic CBI hypothesis.

**Definition 2.4 (**CBI **problem).** A CBI problem is a tuple $\langle \Sigma, s_0 \rangle$ consisting of a CBI setup $\Sigma$ and a new input $s_0 \in \mathcal{S}$. The task is to predict the output $r_0 = \varphi(s_0)$ associated with $s_0$. To this end, the information provided by $\Sigma$ (essentially the similarity measures and the observed cases) is to be exploited against the background of the CBI hypothesis.[42] □

According to Definition 2.4, the heuristic assumption underlying case-based reasoning and, related to this, the use of similarity as a major concept can be seen as the characteristic properties of CBI. In fact, this is what makes a CBI problem a *special* type of prediction problem. In this connection, let us mention that different kinds of knowledge can be distinguished in case-based problem solving:[43] The CBI hypothesis itself can be seen as a kind of (heuristic) *meta-knowledge*, whereas the similarity measures $\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$ often encode *domain-specific knowledge*. The memory of cases, $\mathcal{M}$, corresponds to the experience and represents *empirical* knowledge. Needless to say, a clear separation is generally not possible, especially since the above types of knowledge strongly influence each other.

Let us now introduce an illustrative example from the field of combinatorial optimization to which we shall return occasionally in subsequent sections.

Example 2.5. A *repetitive combinatorial optimization problem* (RCOP) is identified by a class of combinatorial optimization problems, the instances of which appear in only slightly different form. This kind of problem is particularly interesting from a case-based reasoning perspective [238, 239]. As an example of an RCOP let us consider a class of integer linear programs (ILPs)

$$A \times x \geq y, \quad x \times c \longrightarrow \min,$$

where $A$ and $c$ are fixed and only the right-hand side $y$ varies. We define two concrete problems by means of

$$A_1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}, \qquad c_1 = \begin{pmatrix} 3 \\ 2 \\ 4 \\ 1 \\ 4 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 1 & 3 & 0 & -1 & 0 \\ 0 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 & 3 \end{pmatrix}, \quad c_2 = \begin{pmatrix} 2 \\ 1 \\ 3 \\ 1 \\ 6 \end{pmatrix}.$$

In order to obtain corresponding CBI setups $\Sigma_1$ and $\Sigma_2$, we further formalize these examples as follows:

[42] Again, it should be noted that the mapping $\varphi$ is not known, even though formally it is part of the CBI setup.

[43] The idea of *knowledge containers* [313] provides a useful concept for the representation and organization of different types of knowledge in CBR systems.

$$\begin{aligned}
\mathcal{S} &= \{s = (y_1, \ldots, y_5) \,|\, 0 \leq y_1, \ldots, y_5 \leq 6\}, \\
\mathcal{R} &= \mathfrak{N}_0, \\
\sigma_{\mathcal{S}}(s, s') &= \exp\left(-0.1 \sum_{k=1}^{5} |y_k - y'_k|\right), \qquad (2.30) \\
\sigma_{\mathcal{R}}(r, r') &= \exp\left(-0.1\, |r - r'|\right), \\
\varphi(s) &= \min\{x \times c \,|\, x \in \mathfrak{N}_0, A \times x \geq s\}.
\end{aligned}$$

That is, inputs correspond to the right-hand sides $y$, the entries of which are non-negative integers not larger than six. We suppose the set of inputs to be endowed with the uniform probability measure, i.e., the probability of each vector $s \in \mathcal{S}$ is $7^{-5}$. The output associated with an input is the cost $x^* \times c$ of the corresponding optimal solution $x^*$.

Finally, we define two further CBI setups $\Sigma_1^*$ and $\Sigma_2^*$ by modifying $\varphi$, $\mathcal{R}$ and $\sigma_{\mathcal{R}}$ in (2.30) as follows: $\varphi(s)$ corresponds to the optimal solution $x^*$ itself (rather than its cost), $\mathcal{R}$ is the set of all such solutions, and $\sigma_{\mathcal{R}}$ is the same function as $\sigma_{\mathcal{S}}$ in (2.30). □

### 2.4.2 Non-deterministic inference problems

A non-deterministic CBI setup is defined in the same way as a deterministic one except that outputs are not assumed to be uniquely determined by inputs. Instead, an output $\varphi(s)$ is considered as a random variable. There are different motivations for modeling problems within such a non-deterministic setting. The first and most obvious one is the idea that the process which determines the output associated with a certain input is indeed subject to random influences. For example, suppose that an input is characterized by a tuple $(b, w)$, the numbers of black and white balls in an urn and let the outcome correspond to the number of black balls in a random sample of a certain size. More generally, if observations are chosen at random according to a *joint* probability measure $\mu$ over the input-output space $\mathcal{S} \times \mathcal{R}$, as typically assumed in statistics and machine learning, the output $\varphi(s)$ associated with a fixed input $s \in \mathcal{S}$ corresponds to the conditional measure $\mu(\cdot \,|\, s)$.

The second motivation, which seems to be of considerable practical relevance, is related to the *completeness*, *precision*, and *granularity* of information. It leads us to the idea of problems which are *apparently* non-deterministic in the sense that the outcome is principally determined by the input, but where the characterization of the input is incomplete, imprecise, ambiguous, or not detailed enough. A random outcome associated with an input $s$ is then used for characterizing the (subjective) uncertainty concerning the true but unknown output $r = \varphi(s)$.

As an example of incompleteness consider the description of an input which contains missing attribute values.[44] Sensor information of a mobile robot, which does not allow one to determine the position unequivocally, is an example of ambiguity. The following situation is particularly relevant in connection with (deterministic) search algorithms as a problem solving method: Suppose that an input corresponds to the state of some state space. The result (e.g., whether a solution will be found or not) of applying a certain search operator in a certain situation may then also depend on past as well as future decisions. In other words, the result of a single search decision is uniquely determined for a deterministic (even if heuristic) search algorithm, but it is not known at the time of decision making.

A third motivation for modeling CBI problems within a non-deterministic setting is the assumption that observations might be imprecise. In this case, it is actually true that an input $s \in \mathcal{S}$ determines the outcome $r = \varphi(s)$. Yet, the latter cannot be observed exactly. Again, this kind of uncertainty, which occurs frequently in connection with experimental data, can be modeled by means of a probability measure over $\mathcal{R}$.

Non-determinism should not generally be seen as a drawback. Even if it might be avoided, the gain of information provided by a deterministic setting will often not compensate for the expense of a more precise or detailed characterization of inputs. On the contrary, it may sometimes be advantageous to accept a slight increase in uncertainty in order to reduce complexity.[45]

There are different possibilities of approaching non-deterministic CBI problems. In fact, the way in which such problems are treated depends on the interpretation of the CBI hypothesis within the generalized setting. Let us first consider a hypothesis corresponding to the one in the deterministic setting. That is, it draws conclusions about the similarity of outputs, given the similarity of inputs. As before, a case is defined as a tuple $\langle s, x \rangle \in \mathcal{S} \times \mathcal{R}$ consisting of an input and an output, where $x$ is a realization of the random variable $X$ associated with $s$. Observe, however, that the memory $\mathcal{M}$ may now contain cases $\langle s, x \rangle$ and $\langle s, x' \rangle$ such that $x \neq x'$, which is not possible within the deterministic setting. Indeed, $x$ and $x'$ may be rather different even though the input is the same. This becomes obvious if $X$ follows a uniform distribution over a (large) set $\mathcal{R}$ of outputs. Since realizations are obtained by means of independent experiments, it may be very unlikely to obtain similar outputs for similar (or even identical) inputs. Consequently, the original interpretation of the CBI hypothesis seems rather questionable within the non-deterministic setting.

[44] For certain types of problems, missing attribute values can actually not be avoided. In game playing, for example, where the problem is to make an optimal move, the policy of the opponent can be seen as a "missing attribute."

[45] This point of view seems to be more and more accepted in scientific practice and is in accordance with a general change in the attitude toward uncertainty, which is now regarded as a useful concept [352]. The connection between uncertainty and complexity in the modeling of systems is an important subject of current research in systems science [228, 231].

In fact, it seems more reasonable to interpret the CBI hypothesis in such a way that "similar inputs are characterized by similar random outcomes (i.e., probability distributions on $\mathcal{R}$)." A case is then defined as a tuple $\langle s, \mu \rangle \in \mathcal{S} \times \mathcal{P}(\mathcal{R})$, with $\mathcal{P}(\mathcal{R})$ being the class of probability measures over $\mathcal{R}$. Since the outcome associated with an input $s$ is now a random variable $X$ characterized by the measure $\mu$, it is again deterministic. In other words, we might treat the non-deterministic problem as a special case of the deterministic one by simply defining the set $\mathcal{R}$ of outputs as a class of probability measures (endowed with a related similarity measure) over the set of original outcomes. The other way round, the deterministic setting emerges as a special case of a non-deterministic setting by restricting the class of probability measures to the Dirac measures and identifying the latter with corresponding outputs. Thus, we could restrict ourselves to the deterministic setting and generally think of $\mathcal{R}$ as a class of probability measures. For reasons of clarity and notational convenience, however, we will continue to distinguish the two settings.

**Definition 2.6 (non-deterministic** CBI **setup).** A non-deterministic setup is defined as a 6-tuple

$$\Sigma = \left\langle (\mathcal{S}, \mu_{\mathcal{S}}), \mathcal{R}, \varphi, \sigma_{\mathcal{S}}, \sigma_{\mathcal{P}}, \mathcal{M} \right\rangle ,$$

where the set $\mathcal{S}$ of inputs and the set $\mathcal{R}$ of outputs are countable. The mapping $\varphi : \mathcal{S} \longrightarrow \mathcal{P}(\mathcal{R})$ assigns a probability measure $\mu \in \mathcal{P}(\mathcal{R})$ to each input $s \in \mathcal{S}$. This measure defines the (conditional) probability of outcomes given $s$. Moreover, $\sigma_{\mathcal{P}} : \mathcal{P}(\mathcal{R}) \times \mathcal{P}(\mathcal{R}) \longrightarrow [0,1]$ defines a similarity measure over $\mathcal{P}(\mathcal{R})$. □

It deserves mentioning that the definition of a similarity measure over $\mathcal{P}(\mathcal{R})$ is far from being obvious. It might appear reasonable, for instance, to define similarity in terms of the distance of distribution functions,[46] e.g.,

$$\sigma_{\mathcal{P}}(\mu, \mu') \overset{\mathrm{df}}{=} 1 - \sup_{r \in \mathcal{R}} |\mu(r) - \mu'(r)|.$$

This, however, implies $\sigma_{\mathcal{P}}(\mu, \mu') = 0$ if $\mu$ and $\mu'$ are Dirac measures associated with different outcomes, even if these outcomes are very similar. This example might suggest to define the similarity of two measures as the *expected* similarity $\mathbb{E}(\sigma_{\mathcal{R}}(X_1, X_2))$, where $\sigma_{\mathcal{R}}$ is a similarity measure over $\mathcal{R}$. The (independent) random variables $X_1$ and $X_2$ are distributed according to $\mu_1$ and $\mu_2$, respectively. In this case, however, we will generally have $\sigma_{\mathcal{P}}(\mu, \mu) \ll 1$, which is again a rather questionable property (and contrasts the assumption of reflexivity). In fact, there does not seem to exist an approach that will always lead to reasonable results. Rather, the adequacy of a similarity over $\mathcal{P}(\mathcal{R})$ will depend on the respective application. Often, it will make sense to compare certain characteristic properties of probability distributions. If, for instance, an output $r$ signifies a corresponding

[46] Various proposals of distance measures, such as the Kullback-Leibler divergence, have been made in the relevant literature which is often referred to as information geometry.

monetary gain and if the *expected* gain is considered as the most important aspect of an input, it seems reasonable to let $\sigma_{\mathcal{P}}(\mu_1, \mu_2) = 1 - f(|\mathbb{E}(X_1) - \mathbb{E}(X_2)|)$ for some non-decreasing function $f(\cdot)$.

In connection with the first two types of non-determinism the observations related to an input $s \in \mathcal{S}$ will often not be given in the form of complete measures $\mu$. Rather, such observations appear as (precise) outcomes $r \in \mathcal{R}$, which correspond to realizations of the random variable associated with $s$. This raises the question of how to "collect" cases in the form of *input–probability measure* tuples in order to construct a memory $\mathcal{M}$. One possibility, for instance, is to replace cases by "estimated cases" $\langle s, \widehat{\mu} \rangle$, where $\widehat{\mu}$ is an estimation of the measure $\mu$ derived from observations by means of statistical methods (cf. Section 3.2.2). Such aspects become relevant in connection with the idea of *case-based learning*.

Of course, the aforementioned problem (of collecting observations) does not arise in connection with the third type of uncertainty, i.e., the modeling of imprecise observations. A question related to this model, however, concerns the uniqueness of the measure $\mu$ associated with an input $s \in \mathcal{S}$. Two identical experiments, for instance, might lead to different (imprecise) observations. If we assume this to be caused by random influences, we principally obtain a combination of the first and the third source of non-deterministic outputs. Thus, outcomes should be modeled as probabilities over probability measures, i.e., as higher-order probabilities. Of course, we might also suspect that only the two experimental setups (which correspond to inputs) have not been specified precisely enough, which leads to a combination of the second and the third source of non-determinism.[47]

The above examples of imprecision or ambiguity might suggest that not each type of incomplete information is naturally modeled in a probabilistic way. Imprecise observations in experiemental studies, for instance, could as well be considered as "vague data" in the framework of fuzzy sets. Similar remarks also apply to the probabilistic model characterizing the occurrence of inputs. Thus, it seems reasonable to generalize Definition 2.6 correspondingly.

**Definition 2.7 (generalized non-deterministic** CBI **setup).** Let $\mathcal{F}(X)$ denote the class of normalized uncertainty measures over a (countable) set $X$, i.e., measures $\eta : 2^X \longrightarrow [0,1]$ such that $\eta(\emptyset) = 0$, $\eta(X) = 1$, and $\eta(A) \leq \eta(B)$ for all $A \subseteq B \subseteq X$. A generalized non-deterministic CBI setup is defined as a 6-tuple

$$\Sigma = \big\langle (\mathcal{S}, \eta_{\mathcal{S}}), \mathcal{R}, \varphi, \sigma_{\mathcal{S}}, \sigma_{\mathcal{F}}, \mathcal{M} \big\rangle ,$$

where the set $\mathcal{S}$ of inputs and the set $\mathcal{R}$ of outputs are countable. The information about (the occurence/existence/plausibility of) inputs is characterized by the measure $\eta_{\mathcal{S}} \in \mathcal{F}(\mathcal{S})$. Moreover, $\varphi : \mathcal{S} \longrightarrow \mathcal{F}(\mathcal{R})$ assigns a normalized uncertainty measure $\eta \in \mathcal{F}(\mathcal{R})$ to each input $s \in \mathcal{S}$. The function $\sigma_{\mathcal{F}} : \mathcal{F}(\mathcal{R}) \times \mathcal{F}(\mathcal{R}) \longrightarrow [0,1]$ defines a similarity measure over $\mathcal{F}(\mathcal{R})$. □

[47] Needless to say, the distinction between the first two types of non-determinism is often far from being obvious. This distinction does even give rise to fundamental philosophical questions of (non-)determinism.

Remark 2.8. An obvious generalization of a functional relationship $\varphi : \mathcal{S} \longrightarrow \mathcal{R}$ is a relation $\varphi' \subset \mathcal{S} \times \mathcal{R}$. Given an input $s$, the set of possible outcomes is then given by $A_s = \{r \in \mathcal{R} \,|\, (s,r) \in \varphi'\}$. That is, an input $s$ does not identify a unique output but only a subset of $\mathcal{R}$. This case corresponds to a special type of generalized non-deterministic setup, where the measure $\eta_s = \varphi(s)$ is a $\{0,1\}$-valued *possibility measure* [116]: $\eta_s(A) = 1 \Leftrightarrow A \cap A_s \neq \emptyset$. Likewise, $\eta_{\mathcal{S}}$ is a $\{0,1\}$-valued measure such that $\eta_{\mathcal{S}}(S) = 1 \Leftrightarrow \exists\, s \in S \,:\, A_s \neq \emptyset$. This kind of setting will be explored in Chapters 5 and 6. □

### 2.4.3 Formal models of case-based inference

Our comments in previous sections and in Chapter 1 suggest to distinguish between a strong version of the CBI hypothesis, which concludes from the similarity of inputs on the similarity of outputs in a deterministic way, and a weak version, which only concludes on the *likelihood* of outcomes to be similar. As will be seen in subsequent chapters, these interpretations give rise to different types of predictions. Particularly, outcomes will be considered as being either completely possible or completely impossible according to the strong version, which is generally not the case for the weak interpretation.[48] According to the kind of prediction it seems reasonable to distinguish the following types of CBI *models*:

- A *point-estimation* (such as the predicted class label in IBL) of the unknown outcome $\varphi(s_0)$ is derived.
- A *set-valued prediction* which does not further differentiate between possible candidates is derived. The predicted set will generally be assumed to cover the true outcome, at least with a certain degree of probability.
- The prediction of $\varphi(s_0)$ includes a *valuation* of possible outcomes, e.g., a ranking of the outputs or degrees of likelihood associated with individual outcomes. Such a valuation will generally be realized by means of an uncertainty measure over the set of outputs, for example a probability or possibility measure.

To summarize, in connection with the CBI *problem* we distinguish the way an output is *produced*, whereas the CBI *model* refers to the way (a prediction of) this output is *characterized*. When combining the two types of problems discussed in Sections 2.4.1 and 2.4.2, respectively, with the three types of models suggested above, we obtain the types of case-based inference shown in Fig. 2.3.

As can be seen, the most involved situation arises when characterizing the outcome of a non-deterministic problem by means of an uncertainty measure. In fact, this requires uncertainty concepts of higher order, such as higher-order probabilities.

[48] The terms *deterministic* and *non-deterministic* case-based reasoning as used in [99] refer to these properties. It should be noted that the same terms have been introduced with a different meaning in Section 2.4.1 and Section 2.4.2.

| | prediction in the form of | | |
|---|---|---|---|
| problem | outcome | set | measure |
| deterministic | $\widehat{r}_0 \in \mathcal{R}$ | $\mathcal{C} \subseteq \mathcal{R}$ | $\eta \in \mathcal{F}(\mathcal{R})$ |
| non-deterministic | $\widehat{\mu}_0 \in \mathcal{P}(\mathcal{R})$ | $\mathcal{C} \subseteq \mathcal{P}(\mathcal{R})$ | $\eta \in \mathcal{F}(\mathcal{P}(\mathcal{R}))$ |

**Fig. 2.3.** Possible types of case-based inference.

## 2.5 Summary and remarks

### Summary

– Some background information on similarity-based reasoning resp. case-based reasoning (in the broad sense) has been provided, including a brief outline of the most important methods (NN estimation, IBL, CBR) as well as a discussion concerning the formalization of the similarity concept. Since case-based reasoning is strongly related to instance-based learning, some differences between *model-based* and *instance-based* approaches have been pointed out.

– A new approach to similarity evaluation has been outlined in Section 2.3.3. This method makes use of the Choquet integral resp. the Sugeno integral as an aggregation operator. Thus, a global degree of similarity between a pair of objects is derived from (local) similarities of these objects with respect to individual attributes. A main advantage of the two aggregation operators is their ability to take interdependencies between different attributes into account.

– The difference between reasoning at the *system* (*instance*) *level* and reasoning at the *similarity level* has been emphasized in connection with a comparison between CBI and model-based induction. This distinction will be further explored in subsequent chapters.

– We have introduced a formal and rather general framework in which the task of *case-based inference* has been defined as one of predicting the outcome $r_0$ associated with a new input $s_0$. The characteristic property which distinguishes CBI from other prediction methods is the heuristic assumption underlying case-based reasoning and, related to this, the use of similarity as an essential component of the inference process.

– We have distinguished between *deterministic* CBI problems, in which an input determines the associated output in a unique way, and *non-deterministic* problems, in which the outcome is considered as a random variable.

– Finally, an overview of possible approaches to case-based inference has been given. More precisely, we have distinguished models of CBI according to two dimensions, namely the type of *problem* (deterministic or non-deterministic) and the type of *prediction* (point-estimation, subset of possible outcomes, uncertainty measure over the set of outcomes).

## Remarks

– It has already been said that the assumption of independent and identically distributed (i.i.d) data is typical of statistics and machine learning, even though a weakening of this assumption is an important (and of course practically relevant) objective of current research. In fact, the assumption that past observations are to some extent representative of (and hence relevant to) the future seems to be a minimal prerequisite of inductive inference and, hence, a basic requirement of any type of prediction. From this point of view, statistical reasoning lies somewhere in-between induction and deduction. Indeed, statistical inference generally combines an inductive step, namely the estimation of a (probabilistic) model from observed data, with a deductive step, namely the derivation of (probabilistic) statements from that model.

– In CBR, it is common to speak of the similarity between a case (in the memory) and the target problem. This generally means the similarity between the *problem* (input) associated with that case and the target problem (query input), of course. Alternatively, one often speaks of the similarity between a stored case and the new case. Again, what is actually meant is the similarity between the respective problems.

– CBR systems have been granted several advantages, especially in comparison with model-based and rule-based systems. Notably, CBR can simplify the knowledge acquisition task to a certain degree. In fact, knowledge acquisition is basically realized by collecting relevant experiences (cases) which is clearly less difficult than extracting a model or a set of rules. Moreover, CBR systems dispose of a *graceful degradation* of performance, i.e., they are often able to cope with ill-defined or incompletely specified problems. Finally, they improve incrementally over time in a quite natural way, namely by adding experiences in the form of successfully solved cases.

– In CBR, one usually distinguishes between two phases of case retrieval, namely an *initial matching process* in which a set of plausible candidates is retrieved, and a subsequent process in which a best case among these candidates is selected. Case-based inference, as outlined in this chapter, can obviously be seen as a method that supports the initial matching process.

– In some CBR systems the concept of similarity is actually not as important as our previous comments might suggest. Eventually, the problem is that of retrieving a *useful* or *relevant* case from the memory, and similarity might be seen as only one among several aids (indicators) which can guide the search for such cases. In fact, case retrieval can be realized by means of alternative techniques as well, such as inductive or knowledge-guided indexing and structuring of cases. Again, it is of course possible to consider the retrieved case as the one which is most similar by definition. Needless to say, however, this might contrast the intuitive idea of similarity (cf. Section 2.3).

Moreover, a similarity measure is often defined for the set of inputs (problems) but not for the set of outputs (solutions). In fact, such a measure is actually superfluous if one completely concentrates on one solution, namely the solution associated with the most similar input. As will be seen in later chapters, however, a similarity measure over the set of outcomes is an essential prerequisite of modeling the CBI hypothesis, and for taking the aspect of uncertainty of predictions into account. In this connection, it should be mentioned that in practice, a reasonable similarity measure over the set of outputs is often less difficult to define than a measure over the set of inputs.

– The framework of CBI proposed in this chapter is very generic. This is mainly due to the flexibility in defining the concept of a case. The non-deterministic setting even allows for taking uncertain or incomplete information concerning the description of cases into account. Besides, the performance task of prediction is of a rather general nature. As special cases it includes the prediction of numeric and symbolic values as well as several task types considered in the literature on expert systems [302]. According to [142], the fact that an output is a function of a set of observable attributes (the input) is the main characteristic of prediction. This may be contrasted with the task type of *recognition* which assumes a functional relation in the reverse direction, i.e., it assumes the observable attributes to be determined by the output.

  Nevertheless, let us mention that more complex and perhaps less structured domains might call for further generalizations. See [59] for a view of case-based reasoning as *case completion* which even gives up the distinction between a problem and a solution in connection with the representation of a case.

– Replacing a mapping from inputs to outputs by a (more general) functional relation $\varphi : \mathcal{S} \longrightarrow \mathcal{P}(\mathcal{R})$, as we have done in Definition 2.6, does more or less invalidate the aforementioned distinction between the task types of *prediction* and *recognition*. Namely, a functional relation $\varphi^{-1} : \mathcal{R} \longrightarrow \mathcal{P}(\mathcal{S})$ in the reverse direction – but of the same structure – can be obtained via

$$\mu_{S|(R=r)}(s) = \frac{\mu_{R|(S=s)}(r) \cdot \mu_{\mathcal{S}}(s)}{\mu_{\mathcal{R}}(r)}, \tag{2.31}$$

  where $\mu_{\mathcal{R}} = \varphi(\mu_{\mathcal{S}})$ and $\mu_{R|(S=s)} = \varphi(s)$. Of course, this makes CBI applicable to a wider range of problems. A naïve Bayesian classifier may serve as an example. It assumes an unobservable variable Class (the input) to determine the probability distributions of a set of (independent) observable attributes $A_1, \ldots, A_n$ (the output). The task is to predict the value of Class, given the values of the attributes. More generally, consider parametric methods in the context of statistical inference. In connection with the problem of parameter estimation, for instance, the observed data (the output) is determined by the parameter vector, which corresponds to the input and cannot be observed. However, according to (2.31) we may also exchange the role of the data and the parameter vector.

Thus, we simply consider the data as the input and the parameter vector as the (uncertain) output.[49]

- Despite the differences between (model-based) statistical and case-based inference discussed in Section 2.4, it should be mentioned that the concept of similarity plays an important, even though less emphasized, role in (classical) statistical techniques as well. Here are some examples:
  - *Kernel smoothing* techniques such as the *kernel-based estimation* of probability density functions rely on a closeness assumption which is quite comparable to that of instance-based and case-based reasoning methods.
  - *Multidimensional (classical or ordinal) scaling* takes (dis)similarities between individuals as a point of departure and tries to derive a (vector-valued) description of individuals which is compatible with this information, i.e., which preserves the distance between all pairs of individuals.
  - *Cluster analysis* and *mixture decomposition* are techniques used for identifying concentrations or groups of individuals in a space. A group is called a cluster and should combine individuals which are similar in a certain sense. Thus, similarity (or distance) is used as a basic concept for decomposing the sample data. Besides, a preliminary grouping is often the first step of a data analysis. The construction of a histogram, for instance, involves a decomposition of the data into equivalence classes, a special type of similarity relation.
  - The idea of similarity is also somewhat present in *robust statistics*. The latter term refers to inference procedures which perform (more or less) well even if not all assumptions are completely satisfied, i.e., which are robust to departures from these assumptions. One might be interested, for example, in an inference procedure which yields good estimations of a parameter even if the form of the underlying true distribution does slightly deviate from the assumed form of the distribution. Loosely speaking, we wish to obtain similar estimations for similar types of distributions.
  - *Parameterized statistical models* can often be interpreted as encoding the CBI hypothesis in an implicit way. This is simply due to the fact that most parametric models are defined in terms of continuous functions (i.e., the hypothesis space consists of continuous functions). According to the linear regression model (2.28), for instance, the outcomes of similar (i.e., close with respect to the Euclidean metric) inputs have similar (expected) values resp. probability distributions. The smaller (in absolute size) the parameters $\alpha_k$ are, the stronger this property is developed. The assumption of a *smooth* variation of outcomes becomes even more obvious in connection with generalized regression techniques, such as kernel regression and regression based on spline

[49] The inversion of *causes* (parameters of probabilistic data generating processes) and *effects* (observed data) lies at the heart of the Bayesian paradigm [28] and does also provide the basis of other approaches to statistical analysis, notably the *fiducial* approach of FISHER [147]. In fact, (2.31) does formally correspond to BAYES's Theorem, the first inversion of probabilities and a major conceptual step in the history of statistics.

functions [391]. In fact, since these models are very flexible and allow the relationship between the input and the output to vary over time, only the smoothness of this relation remains as a major assumption.

- A kernel function, one of the key concepts in modern *kernel-based learning* methods [335, 339], can often be interpreted as a kind of similarity function. In this research field, a kernel function is formally defined as finitely positive semi-definite function $\kappa : \mathcal{S} \times \mathcal{S} \longrightarrow \mathfrak{R}$.

# 3. Constraint-Based Modeling of Case-Based Inference

In this chapter, we adopt a constraint-based view of the CBI hypothesis, according to which the similarity of inputs imposes a constraint on the similarity of associated outcomes in the form of a lower bound. A related inference mechanism then allows for realizing CBI as a kind of constraint propagation. We also discuss representational issues and algorithms for putting the idea of *learning* within this framework into action. The chapter is organized as follows: Section 3.1 introduces the aforementioned formalization of the CBI hypothesis. A case-based inference scheme which emerges quite naturally from this formalization is proposed in Section 3.2 and further developed in Section 3.3. Case-based learning is discussed in Section 3.4. In Section 3.5, some applications of case-based inference in the context of statistics are outlined. The chapter concludes with a brief summary and some complementary remarks in Section 3.6.

## 3.1 Basic concepts

### 3.1.1 Similarity profiles and hypotheses

Proceeding from the framework introduced in Section 2.4, the system under consideration can be thought of as the triple $(\mathcal{S}, \mathcal{R}, \varphi)$.[1] The (unknown) functional relation $\varphi$ completely determines the structure of this system at the *instance level*, whereas a memory of observed cases provides only partial information. In connection with CBI, we are interested in utilizing the additional information provided by a CBI setup $\Sigma$ for deriving a corresponding characterization of the system at the *similarity level*. This additional information is mainly contained in the similarity measures.

**Definition 3.1 (similarity profile).** Consider a CBI setup $\Sigma$. The function $h_\Sigma : D_\mathcal{S} \longrightarrow [0,1]$ defined by

$$h_\Sigma(x) \overset{\mathrm{df}}{=} \inf_{s,s' \in \mathcal{S},\, \sigma_\mathcal{S}(s,s')=x} \sigma_\mathcal{R}(\varphi(s), \varphi(s'))$$

is called the similarity profile of $\Sigma$. □

[1] This is in agreement with general systems theory, where an abstract system is defined as a relation on a set [228]. It should also be mentioned that this mathematical structure, even though formally very simple, is general enough for modeling any kind of "real" system.

The similarity profile $h_\Sigma$ is the "fingerprint" of the system $(\mathcal{S}, \mathcal{R}, \varphi)$ at the similarity level and (partly) defines the *similarity structure* of the setup $\Sigma$. Just like $\varphi$ determines dependencies at the instance level, $h_\Sigma$ depicts relations between degrees of similarity: Given the similarity of two inputs, it provides a lower bound to the similarity of the respective outcomes. It hence conveys a precise idea of the extent to which the application at hand actually meets the CBI hypothesis, i.e, it can be interpreted as a (multi-dimensional) quantification of the degree to which the CBI hypothesis holds true.[2] In fact, the stronger the similarity structure of $(\mathcal{S}, \mathcal{R}, \varphi)$ is developed, the more constraining the similarity profile will be. Note that the domain and the codomain of $h_\Sigma$ are one-dimensional, whereas $\mathcal{S}$ and $\mathcal{R}$ are generally of higher dimension. Thus, a similarity profile represents knowledge about the system structure $\varphi$ in a *condensed* form. (We will return to the relation between $h_\Sigma$ and $\varphi$ in Section 3.2.)

Needless to say, the similarity profile of a CBI setup will generally be unknown. This leads us to introduce the related concept of a *similarity hypothesis*.

**Definition 3.2 (similarity hypothesis).** A similarity hypothesis is identified by a function $h : [0,1] \longrightarrow [0,1]$ (and similarity measures $\sigma_\mathcal{S}, \sigma_\mathcal{R}$).[3] The intended meaning of the hypothesis $h$ (or, more precisely, the hypothesis $(h, \sigma_\mathcal{S}, \sigma_\mathcal{R})$) is the assumption that

$$\forall\, s, s' \in \mathcal{S} \,:\, (\sigma_\mathcal{S}(s, s') = x) \Rightarrow (\sigma_\mathcal{R}(\varphi(s), \varphi(s')) \geq h(x))\,. \tag{3.1}$$

A hypothesis $h$ is called *stronger* than a hypothesis $h'$ if $h' \leq h$ and $h \not\leq h'$. Let $\Sigma$ be a CBI setup with similarity profile $h_\Sigma$. We say that $\Sigma$ *satisfies* the hypothesis $h$, or that $h$ is *admissible*, if $h(x) \leq h_\Sigma(x)$ for all $x \in D_\mathcal{S}$. □

A similarity hypothesis $h$ is thought of as an approximation of a similarity profile $h_\Sigma$. It thus defines a formal model of the CBI hypothesis for the application at hand, as represented by the setup $\Sigma$. In Section 2.4, it has already been mentioned that different types of hypotheses might be of different expressive power. This remark becomes more obvious now. Since a similarity profile $h_\Sigma$ is a condensed representation of $\varphi$, a similarity hypothesis $h$ will generally be less constraining than a hypothesis which is directly related to $\varphi$, that is, an approximation $\widehat{\varphi} : \mathcal{S} \longrightarrow \mathcal{R}$ of $\varphi$. Yet, a similarity profile has a relatively simple structure which facilitates the formulation, derivation, or adaptation of hypotheses (cf. Section 3.4).

A similarity hypothesis can originate from different sources. Firstly, it might express a purely heuristic quantification of the CBI assumption. In this case, it is often expressed as "*the more* similar two inputs are, *the more* similar the corresponding outputs are." The concept of a similarity profile, as introduced above,

[2] There are obvious ways of deriving a one-dimensional quantification, for example a (weighted) mean of the values $\{h_\Sigma(x) \,|\, x \in D_\mathcal{S}\}$.

[3] Note that is would be sufficient to define a hypothesis on $D_\mathcal{S}$. Quite often, however, it will indeed appear more convenient to let $\mathrm{dom}(h) = [0,1]$, especially if $|D_\mathcal{S}|$ is large. Otherwise, $\mathrm{dom}(h) = [0,1]$ can still be assumed without loss of generality, simply by letting $h(x) = 1$ for all $x \notin D_\mathcal{S}$.

reveals that this kind of formulation implicitly makes a stronger assumption than the simple "similar inputs imply similar outputs" hypothesis. Namely, it suggests the function $h_{\Sigma}$ associated with a setup $\Sigma$ to be increasing, or at least non-decreasing. More precisely, this formulation may be understood as "the more similar two inputs are, the larger is the lower similarity bound of the associated outcomes." Therefore, we call $h$ a *strict hypothesis* if it is a non-decreasing function. Moreover, we say that a setup $\Sigma$ satisfies the CBI hypothesis in the strict sense if $h_{\Sigma}$ is non-decreasing.

Secondly, it is a natural idea to consider the acquisition of hypotheses as a problem of (empirical) *learning*, i.e., to learn hypotheses from observed (pairs of) cases. This way, CBI combines *instance-based learning*, which essentially corresponds to the collection of cases, and *model-based learning*, namely the learning of similarity hypotheses. The assumption that the CBI hypothesis applies in a strict sense serves an (additional) inductive bias in connection with the model-based aspect of learning. In fact, since it suffices to consider non-decreasing functions $h$ as candidates for approximating $h_{\Sigma}$, the hypothesis space $\mathcal{H}$ under consideration is reduced correspondingly.

REMARK 3.3. Observe that the CBI hypothesis can be enforced to hold true in the strict sense by adapting the similarity measure $\sigma_{\mathcal{S}}$ (and, hence, changing the CBI setup correspondingly). In fact, one can always determine a bijective mapping $f : D_{\mathcal{S}} \longrightarrow D_{\mathcal{S}}$ such that $h_{\Sigma}$ is non-decreasing if $\sigma_{\mathcal{S}}$ is replaced by $\sigma'_{\mathcal{S}} = f \circ \sigma_{\mathcal{S}}$. Seen from this perspective, one may always assume that the strict CBI hypothesis is actually valid and simply explain the opposite by the inadequacy of the (originally) chosen similarity measure.[4] □

REMARK 3.4. A strict similarity hypothesis $h$ is closely related to the concept of a *gradual inference rule* in fuzzy set-based approximate reasoning. A gradual rule is a special kind of fuzzy rule of the form "the more $X$ is in $A$, the more $Y$ is in $B$," where $A$ and $B$ are fuzzy sets modeling some gradual concepts. The application of this kind of fuzzy rule in the context of CBI will be discussed in Section 6.1. □

EXAMPLE 3.5. Fig. 3.1 shows the similarity profiles $h_{\Sigma_1}$ and $h_{\Sigma_2}$ of the CBI setups $\Sigma_1$ and $\Sigma_2$ defined by the (repetitive) ILP problems in Example 2.5.[5] As can be seen, these functions are indeed increasing. Moreover, the similarity structure of $\Sigma_1$ is developed more strongly than the structure of $\Sigma_2$. The same remarks apply to the setups $\Sigma_1^*$ and $\Sigma_2^*$, the similarity profiles of which are shown in the same figure. □

[4] Though this would again degrade the CBI hypothesis to a trivial assumption (see the discussion in Section 2.2.3).

[5] We plotted the polygonal line connecting the points $\{(x, h_{\Sigma}(x)) \,|\, x \in D_{\mathcal{S}}\}$.

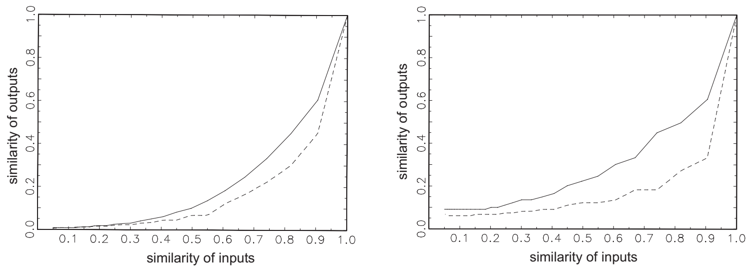


**Fig. 3.1.** Left: Similarity profiles $h_{\Sigma_1}$ (solid line) and $h_{\Sigma_2}$ of the (repetitive) ILP problems defined in Example 2.5. Right: Similarity profiles $h_{\Sigma_1^*}$ (solid line) and $h_{\Sigma_2^*}$ defined in the same example.

EXAMPLE 3.6. Let $(\mathcal{S}, \Delta_{\mathcal{S}})$ and $(\mathcal{R}, \Delta_{\mathcal{R}})$ be metric spaces and suppose $\varphi : \mathcal{S} \longrightarrow \mathcal{R}$ to be Lipschitz continuous, i.e., there is a constant $L > 0$ such that $\Delta_{\mathcal{R}}(\varphi(s), \varphi(s')) \leq L\Delta_{\mathcal{S}}(s, s')$ for all $s, s' \in \mathcal{S}$. Moreover, suppose $\sigma_{\mathcal{S}}$ to be $\Delta_{\mathcal{S}}$-related (via $f$) and $\sigma_{\mathcal{R}}$ to be $\Delta_{\mathcal{R}}$-related (via $g$). Then, $h = g \circ Lf^{-1}$ is an admissible hypothesis for the corresponding CBI setup. □

REMARK 3.7. It has already been suggested in Definition 3.2 to characterize a similarity hypothesis in a more precise way, namely as a triple $(h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. Indeed, the essential aspect in connection with a hypothesis $h$ is the fact that it relates degrees $x$ of the similarity scale $D_{\mathcal{S}}$ (resp. the unit interval) to degrees $y = h(x)$ of the scale $D_{\mathcal{R}}$ (resp. the unit interval). Thus, the meaning of a hypothesis $h$ strongly depends on the similarity functions $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ in the sense that changing these functions would also change the meaning of $h$. Particularly, two hypotheses $h, h'$ as well as the similarity profiles associated with two systems $(\mathcal{S}, \mathcal{R}, \varphi)$ and $(\mathcal{S}, \mathcal{R}, \varphi')$ are not comparable unless the underlying similarity measures are identical. □

### 3.1.2 Generalized similarity profiles

There are two characteristic features of case-based reasoning which are worth mentioning in connection with the concept of a similarity profile and which suggest to generalize Definition 3.1. As will be seen, this generalization makes a similarity profile more suitable for supporting certain (case-based) problem solving strategies.

Firstly, CBI methods do usually not take the complete memory $\mathcal{M}$ of cases into account when solving a new problem. Rather, the attention is drawn to the most similar cases,[6] since less similar cases are assumed to hardly improve the solution (prediction) quality. Indeed, utilizing the complete memory may affect the system

[6] The problem of searching these cases efficiently is closely related to the topics of *case retrieval* and *case indexing* (cf. Section 2.2).

efficiency adversely, at least if the latter does not only take the quality of a solution (prediction) into consideration but also the time which has been spend on deriving it [355, 353]. Secondly, CBI problems might be solved repeatedly by using the same memory $\mathcal{M}$ of cases. One may then benefit from the fact that the memory does not change by adjusting the formalization of the similarity structure to $\mathcal{M}$.

As already announced above, we are now going to introduce some generalizations of Definition 3.1 which are motivated by the two aforementioned aspects.

**Definition 3.8 ($k$-selection).** Let $\mathcal{M} = (\langle s_1, r_1\rangle, \ldots, \langle s_n, r_n\rangle)$, $k \leq n$, and consider an input $s_0 \in \mathcal{S}$. The *extended* $k$-selection $\mathcal{N}_k^{ex}(\mathcal{M}, s_0)$ is defined as a subsequence of $\mathcal{M}$ such that

$$\begin{aligned} \langle s_\jmath, r_\jmath\rangle \in & \mathcal{N}_k^{ex}(\mathcal{M}, s_0) \;\Leftrightarrow \\ & \operatorname{card}\{1 \leq \imath \leq n \,|\, \sigma_{\mathcal{S}}(s_0, s_\jmath) < \sigma_{\mathcal{S}}(s_0, s_\imath)\} < k. \end{aligned}$$

The $k$-selection $\mathcal{N}_k(\mathcal{M}, s_0)$ is defined such that

$$\begin{aligned} \langle s_\jmath, r_\jmath\rangle \in & \mathcal{N}_k(\mathcal{M}, s_0) \;\Leftrightarrow \\ & \operatorname{card}\{1 \leq \imath < \jmath \,|\, \langle s_\imath, r_\imath\rangle \in \mathcal{N}_k^{ex}(\mathcal{M}, s_0)\} < k. \end{aligned}$$

Thus, $\mathcal{N}_k(\mathcal{M}, s_0)$ is exactly of length $k$, whereas $\mathcal{N}_k^{ex}(\mathcal{M}, s_0)$ might consist of more than $k$ cases. □

**Definition 3.9 ($(n, k)$-similarity profile).** Consider a CBI setup $\Sigma$. We define the $(n, k)$-similarity profile

$$h_\Sigma^{(n,k)} : D_{\mathcal{S}} \longrightarrow [0, 1]$$

associated with $\Sigma$ as follows: For all $x \in D_{\mathcal{S}}$, the value $h_\Sigma^{(n,k)}(x)$ is given by the maximal value $y \in [0, 1]$ such that

$$\begin{aligned} \forall\, \mathcal{M} \in \mathcal{M}^n \,\forall\, s_0 \in \mathcal{S} \;\forall\; & \langle s, \varphi(s)\rangle \in \mathcal{N}_k(\mathcal{M}, s_0) \;: \\ & \sigma_{\mathcal{S}}(s, s_0) = x \;\Rightarrow\; \sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)) \geq y, \end{aligned}$$

where $\mathcal{M}^n$ denotes the class of memories of size $n$. □

According to Definition 3.9, the concept of an $(n, k)$-similarity profile corresponds to statements of the following form: "Let $\mathcal{M}$ be an arbitrary memory of size $n$. If two inputs $s_0 \in \mathcal{S}$ and $s \in \mathcal{M}$ are $x$-similar and $s$ is among the inputs in $\mathcal{M}$ which are most similar to $s_0$, then the similarity of the outcomes $\varphi(s_0)$ and $\varphi(s)$ is at least $h_\Sigma^{(n,k)}(x)$." We have $h_\Sigma \leq h_\Sigma^{(n,k)}$ for all $1 \leq k \leq n$, where $n \in \mathfrak{N}$ and $n \leq |\mathcal{S}|$ if $\mathcal{S}$ is finite. This inequality holds due to the fact that $h_\Sigma^{(n,k)}$ is less constrained than $h_\Sigma$, which can be grasped as follows: For $s, s_0 \in \mathcal{S}$ (and $\sigma_{\mathcal{S}}(s, s_0)$ small enough) it might happen that $s \in \mathcal{S}$ is *not relevant* for $s_0$ in the sense that

$$\forall \mathcal{M} \in \mathcal{M}^n : \langle s, \varphi(s) \rangle \notin \mathcal{N}_k(\mathcal{M}, s_0).$$

Now, if neither $s$ is relevant for $s_0$ nor vice versa, the value $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0))$ does no longer constrain the lower bound $h_{\Sigma}^{(n,k)}(\sigma_{\mathcal{S}}(s, s_0))$. Quite often, however, $h_{\Sigma}$ and $h_{\Sigma}^{(n,k)}$ will differ but slightly, at least if $n - k$ is small in relation to the size of the set $\mathcal{S}$.

Remark 3.10. In connection with a "selective" CBI strategy it might be reasonable to require the most similar cases to be (pairwise) different. This amounts to considering only those memories induced by sequences of (pairwise) different inputs. Statistically speaking, a memory $\mathcal{M}$ is then determined by a random sample from $\mathcal{S}$ *without* replacement. Of course, Definition 3.9 can be modified accordingly. □

**Definition 3.11 ($\mathcal{M}$-similarity profile).** Consider a CBI setup $\Sigma$ with memory $\mathcal{M}$. We define $h_{\Sigma}^{\mathcal{M}} : D_{\mathcal{S}} \longrightarrow [0, 1]$ by means of

$$h_{\Sigma}^{\mathcal{M}}(x) \overset{\text{df}}{=} \inf_{s \in \mathcal{M}^{\downarrow}, s_0 \in \mathcal{S}, \sigma_{\mathcal{S}}(s, s_0) = x} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)).$$

This function is called the $\mathcal{M}$-similarity profile of $\Sigma$. □

**Definition 3.12 (($\mathcal{M}, k$)-similarity profile).** Consider a CBI setup $\Sigma$ with memory $\mathcal{M}$. We define $h_{\Sigma}^{(\mathcal{M},k)} : D_{\mathcal{S}} \longrightarrow [0, 1]$ as follows: For all $x \in D_{\mathcal{S}}$, the value $h_{\Sigma}^{(n,k)}(x)$ is given by the maximal value $y \in [0, 1]$ such that

$$\begin{aligned} &\forall s_0 \in \mathcal{S} \; \forall \mathcal{T} \in \mathcal{N}_k(\mathcal{M}, s_0) \; \forall \langle s, r \rangle \in \mathcal{T} : \\ &\qquad (\sigma_{\mathcal{S}}(s, s_0) = x) \Rightarrow (\sigma_{\mathcal{R}}(r, \varphi(s_0)) \geq y) \end{aligned}$$

holds true. The function $h_{\Sigma}^{(n,k)}$ is called the $(\mathcal{M}, k)$-similarity profile of $\Sigma$. □

The above definitions reveal that a $(\cdot, k)$-profile corresponds to the idea of using only $k$ of the stored cases for CBI. Likewise, passing from a similarity profile to an $(\mathcal{M}, \cdot)$-similarity profile is motivated by the idea of repeatedly using a fixed memory $\mathcal{M}$ of cases for solving CBI problems. A profile $h_{\Sigma}^{\mathcal{M}}$, for instance, corresponds to rules of the following form: "Given the memory $\mathcal{M}$ and two $x$-similar inputs $s_0 \in \mathcal{S}$ and $s \in \mathcal{M}$, the similarity of the outcomes $\varphi(s_0)$ and $\varphi(s)$ is at least $h_{\Sigma}^{\mathcal{M}}(x)$." The relations $h_{\Sigma} \leq h_{\Sigma}^{(n,k)} \leq h_{\Sigma}^{\mathcal{M},k}$ and $h_{\Sigma} \leq h_{\Sigma}^{\mathcal{M}} \leq h_{\Sigma}^{\mathcal{M},k}$ hold obviously true for all memories $\mathcal{M}$ and $k \leq n = |\mathcal{M}|$. Passing from a profile $h_{\Sigma}$ to a profile $h_{\Sigma}^{\mathcal{M}}$ will generally have a considerable effect on the quantification of the similarity profile, and the smaller the memory $\mathcal{M}$ is, the stronger this effect will be. In fact, a profile $h_{\Sigma}$ is determined by the similarity relations between *arbitrary* cases $c$ and $c'$, whereas $c$ must be an element of $\mathcal{M}$ in connection with $h_{\Sigma}^{\mathcal{M}}$.

The generalization of Definition 3.2 in accordance with the generalization of similarity profiles is straightforward. We may then speak, e.g., of a similarity hypothesis related to an $\mathcal{M}$-similarity profile or to an $(n,k)$-profile. In subsequent sections of this chapter we will restrict ourselves mainly to the consideration of (ordinary) similarity profiles and related hypotheses, although a further generalization will be introduced in Section 3.3.2. Most often, it will be obvious how to transfer corresponding results.

## 3.2 Constraint-based inference

### 3.2.1 A constraint-based inference scheme

In this section, we shall introduce an inference scheme which emerges quite naturally from the constraint-based view of the CBI hypothesis as formalized in the previous section. Consider a CBI problem $\langle \Sigma, s_0 \rangle$ and suppose that $\Sigma$ satisfies the hypothesis $h$. If the memory $\mathcal{M}$ contains the input $s_0$, i.e., if $\mathcal{M}$ contains a case $\langle s, r \rangle$ such that $s = s_0$, the correct outcome $r_0 = r$ can simply be retrieved from $\mathcal{M}$. Otherwise, we can derive the following restriction:

$$r_0 \in \widehat{\varphi}_{h,\mathcal{M}}(s_0) \overset{\mathrm{df}}{=} \bigcap_{\langle s,r\rangle \in \mathcal{M}} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s_0))}(r), \tag{3.2}$$

where $\widehat{\varphi}_{h,\emptyset}(s_0) \overset{\mathrm{df}}{=} \mathcal{R}$ by convention and the *$\alpha$-neighborhood* of an output $r \in \mathcal{R}$ is defined by the set of all outcomes $r'$ which are at least $\alpha$-similar to $r$:

$$\mathcal{N}_\alpha(r) \overset{\mathrm{df}}{=} \{r' \in \mathcal{R} \,|\, \sigma_{\mathcal{R}}(r,r') \geq \alpha\}. \tag{3.3}$$

Thus, according to the constraint-based interpretation the task of case-based inference can be seen as one of deriving and representing the set (3.2), or an approximation thereof. This may become difficult if, for instance, the definition of the similarity $\sigma_{\mathcal{R}}$ and, hence, the derivation of a neighborhood are complicated. The sets (3.3) may also become large, in which case they cannot be represented by simply enumerating their elements.

In this connection, it should be noted that (3.2) remains correct if the intersection is taken over $k < n$ of the inputs $s \in \mathcal{M}^{\downarrow}$. Since less similar inputs will often hardly contribute to the precision of predictions, it might indeed be reasonable to proceed from $k$ inputs maximally similar to $s_0$, especially if the intersection of neighborhoods (3.3) is computationally complex. Besides, it is worth mentioning that (3.2) can be approached efficiently by means of *parallel computation techniques*. In fact, the sets which have to be combined (via intersection) can be derived independently of each other. Moreover, the (associative) combination itself can be realized in an arbitrary order. Thus, a parallel implementation of

(3.2) is (more or less) straightforward and will enable the exploitation of relatively large memories.

Of course, while assuming the profile of a CBI setup to be unknown, one cannot guarantee the admissibility of a hypothesis $h$ and, hence, the correctness of (3.2). That is, it might happen that $\varphi(s_0) \notin \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. In fact, we might even have $\widehat{\varphi}_{h,\mathcal{M}}(s_0) = \emptyset$. Nevertheless, taking for granted that $h$ is indeed a good approximation of $h_\Sigma$, it seems reasonable to derive $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ according to (3.2) as an approximation of $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s_0)$ (while keeping the hypothetical character of $h$ in mind). This situation reflects the heuristic character of CBI as a problem solving method. Nevertheless, by quantifying the probability of obtaining correct predictions, our results in Section 3.4 will provide a sound basis of this approach.

A similarity profile as well as a similarity hypothesis relate degrees of similarity to one another: Given the similarity of two inputs, they conclude on the similarity of the related outcomes. Thus, the similarity relations between observed cases constitute the principal information from which a case-based inference scheme proceeds. This motivates the following definition.

**Definition 3.13 (similarity structure).** Consider a CBI setup $\Sigma$ with $\mathcal{M}$ being the associated memory (2.29) of cases and let $s_0$ be a new input. The similarity structure of the CBI problem $\langle \Sigma, s_0 \rangle$ is defined by the similarity profile $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ of $\Sigma$ resp. a corresponding hypothesis $(h, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ together with the similarity structure

$$\mathsf{SST}(\mathcal{M}, s_0) \overset{\mathrm{df}}{=} \left\{ z_{\imath\jmath} = (x_{\imath\jmath}, y_{\imath\jmath}) \,|\, 1 \leq \imath < \jmath \leq n \right\} \cup \left\{ x_{0\jmath} \,|\, 1 \leq \jmath \leq n \right\}$$

of the *extended memory* $(\mathcal{M}, s_0)$. Here, the values $x_{\imath\jmath}$ and $y_{\imath\jmath}$ are defined as $x_{\imath\jmath} \overset{\mathrm{df}}{=} \sigma_\mathcal{S}(s_\imath, s_\jmath)$ and $y_{\imath\jmath} \overset{\mathrm{df}}{=} \sigma_\mathcal{R}(r_\imath, r_\jmath)$. We will generally assume the similarity profile $h_\Sigma$ resp. the hypothesis $h$ to be given and simply call $\mathsf{SST}(\mathcal{M}, s_0)$ the similarity structure of $\langle \Sigma, s_0 \rangle$. Moreover, we define the *partial* similarity structure $\mathsf{pSST}(\mathcal{M}, s_0)$ by the set $\{x_{0\jmath} \,|\, 1 \leq \jmath \leq n\}$. □

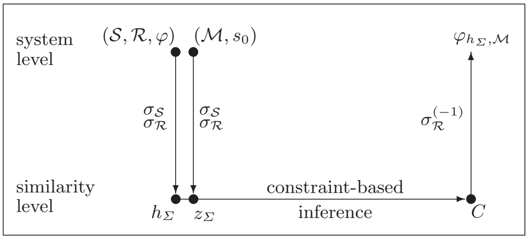

system level
$(\mathcal{S}, \mathcal{R}, \varphi)$
$(\mathcal{M}, s_0)$
$\varphi_{h_\Sigma,\mathcal{M}}$
$\sigma_\mathcal{S}$ $\sigma_\mathcal{R}$
$\sigma_\mathcal{S}$ $\sigma_\mathcal{R}$
$\sigma_\mathcal{R}^{(-1)}$
similarity level
constraint-based inference
$h_\Sigma$
$z_\Sigma$
$C$

**Fig. 3.2.** Illustration of the case-based (similarity-based) inference process.

Even though the inference scheme (3.2) is rather simple, it is worth reconsidering it from an abstract point of view. This will reveal some basic ideas of our approach to CBI, which becomes more involved within the probabilistic setting of Chapter 4. The overall CBI process as illustrated in Fig. 3.2 can be characterized as follows:

– In a first step, the problem $\langle \Sigma, s_0 \rangle$ is characterized at the *similarity level* by means of its *similarity structure*. In fact, $h_\Sigma$ resp. $z_\Sigma = \mathsf{SST}(\mathcal{M}, s_0)$ can be seen as the "image" of the system $(\mathcal{S}, \mathcal{R}, \varphi)$ resp. the (extended) memory $(\mathcal{M}, s_0)$ under the transformation defined by the similarity measures $\sigma_\mathcal{S}$ and $\sigma_\mathcal{R}$. This mapping realizes a projection from an often high-dimensional (and non-numerical) *instance space* $\mathcal{S} \times \mathcal{R}$ into the two-dimensional similarity space $D_\mathcal{S} \times D_\mathcal{R}$, which is usually more accessible to analytical methods. Still, this projection is not (information-)theoretically justified like, say, dimension reduction techniques such as principal component analysis in statistics. Rather, it is guided by the heuristic assumption that the similarity structure of the problem $\langle \Sigma, s_0 \rangle$ represents useful information.

– The main step of the CBI process is then to utilize the similarity structure of the problem for constraining the unknown outcome $r_0$ at the similarity level. The corresponding constraints $C$ are *implicit* in the sense that they are expressed in terms of the (bilateral) concept of similarity, i.e., they do not refer to the output itself.

– Finally, the observed outputs come into play. In conjunction with a transformation $\sigma_\mathcal{R}^{(-1)} : \mathcal{R} \times [0, 1] \longrightarrow 2^\mathcal{R}$, which is inversely related to $\sigma_\mathcal{R}$ via

$$\sigma_\mathcal{R}^{(-1)}(r, \alpha) \overset{\mathrm{df}}{=} \{r' \in \mathcal{R} \,|\, \sigma_\mathcal{R}(r, r') \geq \alpha\}, \tag{3.4}$$

they are used for translating the constraints $C$ at the similarity level into constraints on outcomes at the instance level. According to (3.2), these constraints are combined conjunctively by means of an intersection.

Two characteristics of case-based (similarity-based) inference as introduced above are worth mentioning. Firstly, CBI is *indirect* in the sense that the given information is not used for drawing inferences about the unknown output $r_0$ directly. Rather, it is used for deriving evidence concerning similarity degrees $\sigma_\mathcal{R}(r_0, r_k)$, which are then translated into evidence about outcomes. Secondly, CBI is *local* in the sense that the rules (3.1) associated with a hypothesis $h$ derive evidence concerning the value $r_0$ from single cases. These pieces of evidence have still to be combined in order to obtain the constraint implied by the complete memory $\mathcal{M}$. Within the deterministic framework of this chapter, the combination of evidence derived from different cases is accomplished by (3.2), i.e., by means of a simple intersection of sets. As will be seen in Chapter 4, this problem becomes more complicated within a probabilistic setting.

Needless to say, the stronger the similarity structure of a setup $\Sigma$ is developed, the more successful CBI will be. Within our framework, we have quantified the degree to which the CBI hypothesis holds true for the setup $\Sigma$ by means of the similarity profile $h_\Sigma$. This quantification, however, may appear rather restrictive. In fact, the derivation of valid predictions according to (3.2) necessitates the use of lower similarity bounds, which leads to a kind of worst case analysis. The existence of some "exceptional" pairs of cases, for instance, might call for small values $h_\Sigma(x)$ of the similarity profile $h_\Sigma$. Consequently, the predictions (3.2) which reflect the success of the CBI process (cf. Section 3.4) might become imprecise even though the similarity structure of $\Sigma$ is otherwise strongly developed. This observation serves as a main motivation for the consideration of *local* similarity profiles in Section 3.3.2 and for the probabilistic generalization of the constraint-based approach which we will turn to in Chapter 4.

From a mathematical point of view, the decisive aspect of the inference scheme in Fig. 3.2 is the fact that it is based on the analysis, not of the original data, but of *transformed data* which depicts a certain *relation* between original observations. Considering these observations in pairs, the original data (represented by the memory $\mathcal{M} \subseteq \mathcal{S} \times \mathcal{R}$) is transformed into the new set of data

$$\left\{ (\sigma_{\mathcal{S}}(s, s'), \sigma_{\mathcal{R}}(r, r')) \,|\, \langle s, r \rangle, \langle s', r' \rangle \in \mathcal{M} \right\} . \tag{3.5}$$

As opposed to functional relations related to the instance level, which are mappings of the form $\mathcal{S} \longrightarrow \mathcal{R}$, the result $h$ of the analysis of (3.5) provides information about the *relation* $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s'))$ between outcomes $\varphi(s), \varphi(s')$, given the *relation* $\sigma_{\mathcal{S}}(s, s')$ between inputs $s$ and $s'$. Then, given an observation $\langle s, r \rangle$ and a new input $s_0$ and, hence, the relation $\sigma_{\mathcal{S}}(s, s_0)$, $h$ is used for specifying the relation $\sigma_{\mathcal{R}}(r, r_0)$ between $r$ and $r_0 = \varphi(s_0)$. Finally, the inverse transformation $\sigma_{\mathcal{R}}^{(-1)}$ is used for translating information about $r$ and $\sigma_{\mathcal{R}}(r, r_0)$ into information about $r_0$ itself. Moreover, the *combination of evidence* concerning $r_0$ becomes necessary if this kind of information has been derived from different observations $\langle s_1, r_1 \rangle, \ldots, \langle s_n, r_n \rangle$.

In our case, the relation between observations corresponds to their similarity, the function $h$ defines an (estimated) lower bound in the form of (an approximation of) the similarity profile, and the combination of evidence is realized by the intersection of individual predictions. This, however, is by no means compulsory. Indeed, one might think of basing inference procedures on alternative specifications, such as the differences $\sigma_{\mathcal{S}}(s, s') = s - s'$ and $\sigma_{\mathcal{R}}(r, r') = r - r'$.[7] Then, a least squares approximation $h$ of the transformed data provides an estimation of the difference between two outcomes, given the difference between the respective inputs. Examples of this kind of inference can, e.g., be found in economic analysis where a functional relation is often assumed, not between the economic quantities themselves, but between the (temporal) *change* of these quantities. Economic time series $(x_1, \ldots, x_T)$, for instance, are often analyzed in terms of (first-order) differences $\Delta t_k = t_{k+1} - t_k$. Likewise, in preference analysis, a frequently encountered problem is to induce an absolute rating of given entities (in terms of utility

[7] In this example, $\mathcal{S}$ and $\mathcal{R}$ are assumed to be numerical, of course.

degrees) based on pairwise comparisons expressing to what extent one object is preferred to a second one.

Remark 3.14. The non-deterministic setting of Section 2.4.2 takes account of the fact that an input $s \in \mathcal{S}$ does not determine a unique outcome or that observed outputs might be imprecise. A respective generalization of the inference scheme based on (3.2) will be discussed in Section 3.2.2 below. Simple types of imprecision, however, can also be incorporated directly into (3.2). Suppose for instance, that an output cannot be observed exactly but only up to a certain (similarity) degree $\alpha$ of precision. That is, an observed case $\langle s, r\rangle$ does not imply $\varphi(s) = r$ but only $\varphi(s) \in \mathcal{N}_{1-\alpha}(r)$. Moreover, suppose that $\sigma_{\mathcal{R}}$ is $\top$-transitive, i.e., $\top(\sigma_{\mathcal{R}}(r, r'), \sigma_{\mathcal{R}}(r', r'')) \leq \sigma_{\mathcal{R}}(r', r'')$ for all $r, r', r'' \in \mathcal{R}$ (cf. Section 2.3). We then obtain

$$\varphi(s_0) \in \bigcap_{\langle s,r\rangle \in \mathcal{M}} \mathcal{N}_{\top(h_\Sigma(\sigma_{\mathcal{S}}(s,s_0)),1-\alpha)}(r), \tag{3.6}$$

for all $s_0 \in \mathcal{S}$ as a valid generalization of (3.2). Observe that (3.6) might be interesting in connection with non-deterministic CBI problems, namely when having to use “estimated cases” $\langle s, \widehat{\mu}\rangle$ due to the problem that the true measure $\mu$ might not be observable (cf. Section 2.4.2). In fact, this inference scheme can be applied if a minimal similarity between the true measure $\mu$ and the estimation $\widehat{\mu}$ is guaranteed. □

### 3.2.2 Non-deterministic problems

Within the non-deterministic setting of Section 2.4.2, a similarity profile $h_\Sigma$ of a setup $\Sigma$ is defined by replacing the similarity measure over outputs, $\sigma_{\mathcal{R}}$, by a similarity measure over probability distributions, $\sigma_{\mathcal{P}}$:

$$h_\Sigma : D_{\mathcal{S}} \longrightarrow [0, 1]\,,\; x \mapsto \inf_{s,s' \in \mathcal{S},\, \sigma_{\mathcal{S}}(s,s')=x} \sigma_{\mathcal{P}}(\varphi(s), \varphi(s')).$$

Then, a similarity hypothesis $h$ corresponds to the assumption that

$$\forall\, s, s' \in \mathcal{S} \,:\, \sigma_{\mathcal{S}}(s, s') = x \Rightarrow \sigma_{\mathcal{P}}(\varphi(s), \varphi(s')) \geq h(x)$$

holds true for all $x \in [0, 1]$. Given a memory $\mathcal{M}$ of cases $\langle s_k, \mu_k\rangle$ $(1 \leq k \leq n)$, the inference scheme (3.2) presents itself in the form

$$\mu_0 \in \widehat{\varphi}_{h,\mathcal{M}}(s_0) \overset{\text{df}}{=} \bigcap_{\langle s,\mu\rangle \in \mathcal{M}} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s_0))}(\mu), \tag{3.7}$$

where $\mu_0$ is the probability measure associated with the new input $s_0$ and $\mathcal{N}_\alpha(\mu) \overset{\text{df}}{=} \{\mu' \in \mathcal{P}(\mathcal{R}) \,|\, \sigma_{\mathcal{P}}(\mu, \mu') \geq \alpha\}$ for $\mu \in \mathcal{P}(\mathcal{R})$ and $0 \leq \alpha \leq 1$. Thus, the set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ now defines a class of probability measures, namely the measures which are considered as being possible in connection with the unknown measure $\mu_0$.

**Upper and lower probability bounds.** For a memory $\mathcal{M}$ and a new input $s_0 \in \mathcal{S}$, the set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ as defined in (3.7) corresponds to a *set* of probability measures. Instead of the inference result $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ itself, which might have a relatively complicated structure, one might be interested in the lower and upper probability of *individual outputs* $r \in \mathcal{R}$ according to this set, i.e.

$$\mu_0^{\downarrow}(r) = \min_{\mu \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \mu(r) \quad \text{and} \quad \mu_0^{\uparrow}(r) = \max_{\mu \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \mu(r). \tag{3.8}$$

Let us, therefore, consider a particular (but still reasonable) choice of the similarity $\sigma_{\mathcal{P}}$ which supports an efficient derivation of these probability bounds:

$$\sigma_{\mathcal{P}}(\mu, \mu') \overset{\text{df}}{=} 1 - f\left( \max_{r \in \mathcal{R}} |\mu(r) - \mu'(r)| \right) \tag{3.9}$$

for all $\mu, \mu' \in \mathcal{P}(\mathcal{R})$, where $f : [0,1] \longrightarrow [0,1]$ is (strictly) increasing.[8] The constraint on $\mu_0$ induced by the $k$-th case $\langle s_k, \mu_k \rangle$ is now given in the form of an *interval probability* $[\mu_{0k}^l, \mu_{0k}^u]$, where

$$\mu_{0k}^l(r) = \max\left\{\mu_k(r) - f^{-1}(1 - \sigma_{\mathcal{S}}(s_0, s_k)), 0\right\}, \tag{3.10}$$

$$\mu_{0k}^u(r) = \min\left\{\mu_k(r) + f^{-1}(1 - \sigma_{\mathcal{S}}(s_0, s_k)), 1\right\}, \tag{3.11}$$

and

$$[\mu_{0k}^l, \mu_{0k}^u] \overset{\text{df}}{=} \{\mu \in \mathcal{P}(\mathcal{R}) \,|\, \forall\, r \in \mathcal{R} \,:\, \mu_{0k}^l(r) \le \mu(r) \le \mu_{0k}^u(r)\}. \tag{3.12}$$

Suppose $\widehat{\varphi}_{h,\mathcal{M}}(s_0) \neq \emptyset$ for the overall constraint (3.7). The latter is then also an interval probability:

$$\widehat{\varphi}_{h,\mathcal{M}}(s_0) = [\mu_0^l, \mu_0^u], \tag{3.13}$$

where

$$\mu_0^l(r) = \max_{1 \le k \le n} \mu_{0k}^l(r), \quad \mu_0^u(r) = \min_{1 \le k \le n} \mu_{0k}^u(r) \tag{3.14}$$

for all $r \in \mathcal{R}$. It deserves mentioning that the representation of an interval probability in the sense of (3.12) is not unique. In general, it is possible to represent a given class of probability measures $\mu$ over a set $X$ by means of different intervals $[\mu^l, \mu^u]$ (i.e., lower and upper envelopes $\mu^l : X \longrightarrow [0,1]$ and $\mu^u : X \longrightarrow [0,1]$ such that $\mu^l \le \mu^u$). In fact, the intervals $[\mu_0^l(r_1), \mu_0^u(r_1)]$ are not necessarily minimal, i.e., the lower and upper bounds (3.14) do not necessarily correspond to the optimal bounds (3.8). That is, it might be possible that $\mu_0^l(r_1) < \mu_0^{\downarrow}(r_1)$ or $\mu_0^{\uparrow}(r_1) < \mu_0^u(r_1)$ and, hence, that one can increase $\mu_0^l(r_1)$ or reduce $\mu_0^u(r_1)$ for some $r_1 \in \mathcal{R}$ without changing the associated class (3.13) of probability measures. In other words, it might happen that $\mu_0^l(r_1)$ (resp. $\mu_0^u(r_1)$) is actually not attained by any measure $\mu \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. In the case of finite $\mathcal{R}$, the optimal individual bounds $\mu_0^{\downarrow}(r_1)$ and $\mu_0^{\uparrow}(r_1)$ can be found by solving two simple linear programming problems:

[8] The maximum in (3.9) obviously exists.

$$\text{minimize (maximize) } \mu_0(r_1) \quad \text{s.t.} \quad \begin{cases} \mu_0^l(r) \le \mu_0(r) \le \mu_0^u(r) & (r \in \mathcal{R}) \\ \mu_0(r) \ge 0 \quad (r \in \mathcal{R}) \\ \sum_{r \in \mathcal{R}} \mu_0(r) = 1 \end{cases}$$

REMARK 3.15. The bounds (3.10) and (3.11) associated with a single constraint are already optimal. This can be seen as follows. Let $\alpha_0 = f^{-1}(1 - \sigma_{\mathcal{S}}(s_0, s_k))$ and $\alpha_1 = \min\{\mu_k(r_1), \alpha_0\}$ for some $r_1 \in \mathcal{R}$. That is, $\mu_{0k}^l(r_1) = \mu_k(r_1) - \alpha_1$. If $\alpha_1 = \alpha_0$ then $\mu_k(r_1) \ge \alpha_0$, i.e., there is some $r_2 \in \mathcal{R}$ such that $\mu_k(r_2) \le 1 - \alpha_0$. The probability measure $\mu$ defined by

$$\mu(r) = \begin{cases} \mu_k(r) - \alpha_0 & \text{if} \quad r = r_1 \\ \mu_k(r) + \alpha_0 & \text{if} \quad r = r_2 \\ \mu_k(r) & \text{if} \quad r_1 \ne r \ne r_2 \end{cases}$$

is then an element of $[\mu_{0k}^l, \mu_{0k}^u]$, i.e., the lower bound $\mu_{0k}^l(r_1) = \mu_k(r_1) - \alpha_0$ is indeed attained. Now, suppose $\alpha_1 < \alpha_0$ which means that $\mu_{0k}^l(r_1) = 0$. Since $\mu_k(r_1) = 1 - \sum_{r_1 \ne r \in \mathcal{R}} \mu_k(r)$ and $\mu_k(r_1) < \alpha_0$ it is obviously possible to distribute the probability mass $\alpha_1 = \mu_k(r_1)$ over the elements $r \ne r_1$ such that the measure $\mu$ defined by

$$\mu(r) = \begin{cases} 0 & \text{if} \quad r = r_1 \\ \mu_k(r) + \alpha(r) & \text{if} \quad r \ne r_1 \end{cases}$$

for all $r \in \mathcal{R}$ is an element of the class $[\mu_{0k}^l, \mu_{0k}^u]$, where $\alpha(r) \ge 0$ and $\sum_{r_1 \ne r \in \mathcal{R}} \alpha(r) = \alpha_1$. Thus, the lower bound $\mu_{k0}^l(r_1) = 0$ is again attained. Analogously it is shown that the upper bound $\mu_{k0}^u(r_1)$ is always attained. □

**A Maximum Likelihood approach.** In Section 2.4.2, we have pointed out that it might not be possible to observe the probability measure $\mu$ associated with an input $s$. Rather, a case is often given in the form of a tuple $\langle s, x \rangle$, where $x$ has been chosen at random according to $\mu$. We shall now consider a framework which allows for deriving estimated cases $\langle s, \widehat{\mu} \rangle$ by means of a MAXIMUM LIKELIHOOD (ML) approach.

Let $\mathcal{P}(\mathcal{R})$ consist of a class of parameterized probability measures $\mu_\theta$ $(\theta \in \Theta)$ and suppose that $\sigma_{\mathcal{P}} : \mathcal{P}(\mathcal{R}) \times \mathcal{P}(\mathcal{R}) \longrightarrow [0,1]$ can be expressed as a function of parameter vectors, i.e., the similarity $\sigma_{\mathcal{P}}(\mu_\theta, \mu_{\theta'})$ can be written in terms of the parameter vectors $\theta$ and $\theta'$ for all $\theta, \theta' \in \Theta$. Thus, we can associate a parameter $\theta$ with each input $s$. By thinking of the parameter as an output, we can also identify $\Theta$ by the set of outputs, $\mathcal{R}$, and write $\sigma_{\mathcal{P}}(\mu_\theta, \mu_{\theta'}) = \sigma_{\mathcal{R}}(\theta, \theta')$.[9]

Now, consider a non-deterministic CBI problem. Suppose that $n$ cases $\langle s_k, x_k \rangle$ $(1 \le k \le n)$ have been observed. A reasonable approach to estimating the probability measures $\mu_k$ associated with the inputs $s_k$ is to maximize the likelihood function

[9] Observe that this assumption does not exclude (3.9).

$$\lambda : (\theta_1, \ldots, \theta_n) \mapsto \prod_{1 \leq k \leq n} \mu_{\theta_k}(x_k)$$

subject to the constraints

$$\forall\, 1 \leq \imath, \jmath \leq n \,:\, \sigma_{\mathcal{R}}(\theta_\imath, \theta_\jmath) \geq \sigma_{\mathcal{S}}(s_\imath, s_\jmath).$$

That is, we let $\widehat{\mu}_k = \mu_{\widehat{\theta}_k}$, where the parameter vectors $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$ denote the (constrained) ML estimations. The measure $\mu_0$ associated with a new input $s_0$ is then estimated according to

$$\mu_0 \in \bigcap_{1 \leq k \leq n} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s_0, s_k))}(\widehat{\mu}_k).$$

## 3.3 Case-based approximation

Suppose a hypothesis $h$ (with associated similarity functions $\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$) and a memory $\mathcal{M}$ to be given. By applying (3.2) to all $s \in \mathcal{S}$ (not only to one input $s_0 \in \mathcal{S}$), we obtain a set-valued mapping $\widehat{\varphi}_{h,\mathcal{M}} : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$:[10]

$$\widehat{\varphi}_{h,\mathcal{M}} : s \mapsto \bigcap_{\langle s', r' \rangle \in \mathcal{M}} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s'))}(r'). \tag{3.15}$$

It is readily shown that $\widehat{\varphi}_{h,\mathcal{M}}$ defines an outer approximation of $\varphi$ in the sense that $\varphi(s) \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$ if the hypothesis $h$ is admissible. The mapping $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$, induced by the similarity structure of a CBI setup, can be seen as a simplified but imprecise representation of the system structure $\varphi$. We call $\widehat{\varphi}_{h,\mathcal{M}}$ a *case-based approximation* (CBA) of $\varphi$. Clearly, the stronger the (admissible) hypothesis $h$ is, the more precise the approximation $\widehat{\varphi}_{h,\mathcal{M}}$ becomes. The CBA obtained for the similarity profile $h_\Sigma$, $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$, is the smallest outer approximation of $\varphi$ in the sense that $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h,\mathcal{M}}(s)$ holds true for all $s \in \mathcal{S}$ and admissible hypotheses $h$.

Remark 3.16. Definition (3.15) is not exactly in agreement with our CBI approach in the sense that we may have $\widehat{\varphi}_{h,\mathcal{M}}(s) \neq \{r\}$ for some case $\langle s, r \rangle \in \mathcal{M}$. That is, the prediction $\widehat{\varphi}_{h,\mathcal{M}}(s)$ might contain additional outcomes even though the output $r$ could be retrieved from the memory. It can easily be verified, however, that $\widehat{\varphi}_{h,\mathcal{M}}(s) = \{r\}$ is guaranteed if both measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ are separating. Clearly, a further way of ensuring $\widehat{\varphi}_{h,\mathcal{M}}(s) = \{r\}$ is to modify the definition of a case-based approximation as follows: $\widehat{\varphi}_{h,\mathcal{M}}(s)$ is determined according to (3.15) only if $s \notin \mathcal{M}^{\downarrow}$, otherwise the output is retrieved from $\mathcal{M}$ and is hence given by $\{\varphi(s)\}$. □

[10] This mapping corresponds in some way to what is called the *extensional concept description* in instance-based learning [11].

Let us again mention that (3.2) resp. (3.15) are easily generalized such that only $k < n$ of the most similar cases (represented by a sub-memory $\mathcal{M}' \subseteq \mathcal{M}$) are used for constraining the outcome. Then, we can define an approximation $\widehat{\varphi}_{h,\mathcal{M},k} : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$ by means of

$$\widehat{\varphi}_{h,\mathcal{M},k} : s \mapsto \bigcap_{\langle s',r' \rangle \in \mathcal{T}(s)} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s,s'))}(r'), \tag{3.16}$$

where $\mathcal{T}(s) \overset{\text{df}}{=} \mathcal{N}_k(\mathcal{M}, s)$ or $\mathcal{T}(s) \overset{\text{df}}{=} \mathcal{N}_k^{ex}(\mathcal{M}, s)$.

### 3.3.1 Properties of case-based approximation

It deserves mentioning that the similarity measures principally play the role of *ordinal* concepts within our approach.[11] According to (3.2), the set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ depends only on the relative order of similarity degrees, as specified by the hypothesis $h$ (cf. Remark 3.7). In other words, the sets $D_{\mathcal{S}}$ and $D_{\mathcal{R}}$ can be interpreted as linearly ordered scales of similarity for which only the ordering of the grades of similarity is important. In fact, the numerical encoding is just a matter of convenience and the interval $[0, 1]$ could be replaced by any other linearly ordered scale. In fact, the inference scheme (3.2) can even be generalized in a straightforward way to similarity measures which are defined on a (complete) lattice structure [56, 283].

In order to make the ordinal character of similarity more explicit let us call two similarity measures $\sigma$ and $\sigma'$ (defined over a set $A$) *coherent* if

$$\sigma(a, b) \leq \sigma(c, d) \Leftrightarrow \sigma'(a, b) \leq \sigma'(c, d) \tag{3.17}$$

holds true for all $a, b, c, d \in A$. This definition is in accordance with the relational approach to similarity discussed in Section 2.3 (coherent similarity measures induce the same relation $R$).

**Lemma 3.17.** Let $\sigma : A \times A \longrightarrow [0, 1]$ and $\sigma' : A \times A \longrightarrow [0, 1]$ be coherent similarity measures and let $X = \{\sigma(a, b) \,|\, a, b \in A\}$. Then, a strictly increasing function $f : X \longrightarrow [0, 1]$ exists such that $\sigma' = f \circ \sigma$. □

**Proof.** For $a, b \in A$, let $x = \sigma(a, b)$ and define $f(x) = \sigma'(a, b)$. Obviously, $f$ is well-defined, since the coherency of $\sigma$ and $\sigma'$ implies

$$\sigma(a, b) = \sigma(c, d) \Leftrightarrow \sigma'(a, b) = \sigma'(c, d) \tag{3.18}$$

for all $a, b, c, d \in A$. Moreover, $f$ is strictly increasing, since (3.18) remains valid when replacing the equality relation by the <-relation. □

[11] This should be regarded as a reasonable property. Indeed, considering similarity as a cardinal concept complicates its formalization and raises some difficult semantical questions.

**Proposition 3.18.** Consider a system $(\mathcal{S}, \mathcal{R}, \varphi)$ and a memory $\mathcal{M}$ of cases and let $\sigma_\mathcal{S}$ and $\sigma'_\mathcal{S}$ resp. $\sigma_\mathcal{R}$ and $\sigma'_\mathcal{R}$ be coherent similarity measures. Moreover, denote by $h_\Sigma$ resp. $h'_\Sigma$ the similarity profiles induced by these measures and let $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$ resp. $\widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$ be the case-based approximations defined by $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ resp. $(h'_\Sigma, \sigma'_\mathcal{S}, \sigma'_\mathcal{R})$ via (3.15). We then have $\widehat{\varphi}_{h_\Sigma,\mathcal{M}} = \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$. □

**Proof.** According to Lemma 3.17 there are strictly increasing functions $f$ and $g$ such that $\sigma'_\mathcal{S} = f \circ \sigma_\mathcal{S}$ and $\sigma'_\mathcal{R} = g \circ \sigma_\mathcal{R}$. From (3.18) and $f(\sigma_\mathcal{R}(r, r')) \leq g(\sigma_\mathcal{S}(s, s')) \Leftrightarrow \sigma'_\mathcal{R}(r, r') \leq \sigma'_\mathcal{S}(s, s')$ for all $s, s' \in \mathcal{S}$ and $r, r' \in \mathcal{R}$ then follows that $h'_\Sigma \circ f = g \circ h_\Sigma$. Now, consider $s, s' \in \mathcal{S}$, $r, r' \in \mathcal{R}$ and suppose that $(h_\Sigma \circ \sigma_\mathcal{S})(s, s') \leq \sigma_\mathcal{R}(r, r')$. It follows that

$$\begin{aligned}
\sigma'_\mathcal{R}(r, r') &= (g \circ \sigma_\mathcal{R})(r, r') \\
&\geq (g \circ h_\Sigma \circ \sigma_\mathcal{S})(s, s') \\
&= (h'_\Sigma \circ f \circ \sigma_\mathcal{S})(s, s') \\
&= (h'_\Sigma \circ \sigma'_\mathcal{S})(s, s').
\end{aligned}$$

In the same way it is shown that $(h'_\Sigma \circ \sigma'_\mathcal{S})(s, s') \leq \sigma'_\mathcal{R}(r, r')$ implies $(h_\Sigma \circ \sigma_\mathcal{S})(s, s') \leq \sigma_\mathcal{R}(r, r')$. Consequently, we have $\mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s'))}(r) = \mathcal{N}_{h'_\Sigma(\sigma'_\mathcal{S}(s,s'))}(r)$ for all $s, s' \in \mathcal{S}$, $r \in \mathcal{R}$ and, hence, $\widehat{\varphi}_{h_\Sigma,\mathcal{M}} = \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$. □

In Section 2.4, it was already pointed out that similarity measures might be more or less "discriminating." We are now in the position to put this into more precise terms. Let us call a similarity measure $\sigma$ a *refinement* of a measure $\sigma'$ if $\sigma' = f \circ \sigma$, where $f$ is non-decreasing (i.e., order-preserving) but not (strictly) increasing. Loosely speaking, the measure $\sigma$ uses a richer similarity scale which includes more degrees of similarity, that is $\mathrm{rg}(\sigma') \subsetneq \mathrm{rg}(\sigma)$.

**Proposition 3.19.** Consider a system $(\mathcal{S}, \mathcal{R}, \varphi)$ and a memory $\mathcal{M}$ of cases. Let $\sigma_\mathcal{S}$ be a refinement of $\sigma'_\mathcal{S}$ and $\sigma_\mathcal{R}$ a refinement of $\sigma'_\mathcal{R}$. Moreover, denote by $h_\Sigma$ resp. $h'_\Sigma$ the similarity profiles induced by these measures and let $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}$ resp. $\widehat{\varphi}_{h'_\Sigma,\mathcal{M}}$ be the case-based approximations defined by $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ resp. $(h'_\Sigma, \sigma'_\mathcal{S}, \sigma'_\mathcal{R})$ via (3.15). Then, $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$. □

**Proof.** Consider values $s, s' \in \mathcal{S}$ and $r, r' \in \mathcal{R}$. Suppose that $r' \in \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s'))}(r)$, i.e., $\sigma_\mathcal{R}(r, r') \geq h_\Sigma(\sigma_\mathcal{S}(s, s'))$. Thus, we find $t, t' \in \mathcal{S}$ such that $\sigma_\mathcal{R}(r, r') \geq \sigma_\mathcal{R}(\varphi(t), \varphi(t'))$ and $\sigma_\mathcal{S}(s, s') = \sigma_\mathcal{S}(t, t')$. Since $\sigma'_\mathcal{S} = f \circ \sigma_\mathcal{S}$ and $\sigma'_\mathcal{R} = g \circ \sigma_\mathcal{R}$ for non-decreasing functions $f, g$, we have $\sigma'_\mathcal{S}(s, s') = \sigma'_\mathcal{S}(t, t')$ and $\sigma'_\mathcal{R}(r, r') \geq \sigma'_\mathcal{R}(\varphi(t), \varphi(t'))$. Therefore,

$$h'_\Sigma(\sigma'_\mathcal{S}(s, s')) \leq \sigma'_\mathcal{R}(\varphi(t), \varphi(t')) \leq \sigma'_\mathcal{R}(r, r')$$

and, hence, $r' \in \mathcal{N}_{h'_\Sigma(\sigma_\mathcal{S}(s,s'))}(r)$. □

Of course, generally we will not only have $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}(s)$, as guaranteed by Proposition 3.19, but also $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \neq \widehat{\varphi}_{h'_\Sigma,\mathcal{M}}(s)$ for some $s \in \mathcal{S}$. As an obvious example consider the "least discriminating" case where $g \equiv 1$ on $D_\mathcal{R}$ and, hence, $\sigma'_\mathcal{R} \equiv 1$ on $\mathcal{R} \times \mathcal{R}$, which leads to the trivial prediction $\widehat{\varphi}_{h'_\Sigma,\mathcal{M}} \equiv \mathcal{R}$ on $\mathcal{S}$.

For us to be able to study the approximation capability of (3.15) more thoroughly the system $(\mathcal{S}, \mathcal{R}, \varphi)$ must have a structure which allows us to quantify the quality of a case-based approximation. To this end, let us endow $\mathcal{S}$ and $\mathcal{R}$ with a metric, i.e., let $(\mathcal{S}, \Delta_\mathcal{S})$ and $(\mathcal{R}, \Delta_\mathcal{R})$ be metric spaces. Clearly, a good approximation of $\varphi$ can only be expected if the similarity measures $\sigma_\mathcal{S}$ and $\sigma_\mathcal{R}$ are related to the distance measures $\Delta_\mathcal{S}$ and $\Delta_\mathcal{R}$. We can prove the following result.

**Proposition 3.20.** Suppose that $\sigma_\mathcal{S} = f \circ \Delta_\mathcal{S}$ and $\sigma_\mathcal{R} = g \circ \Delta_\mathcal{R}$ with strictly decreasing functions $f$ and $g$, and

$$\exists\, \varepsilon > 0\, \exists\, \mathcal{S}' \subseteq \mathcal{S} \,:\, \operatorname{card}(\mathcal{S}') < \infty \,\wedge\, \mathcal{S} = \bigcup_{s \in \mathcal{S}'} \bar{\mathfrak{B}}_\varepsilon(s), \tag{3.19}$$

where $\bar{\mathfrak{B}}_\varepsilon(s) \overset{\text{df}}{=} \{s' \in \mathcal{S} \,|\, \Delta_\mathcal{S}(s, s') \le \varepsilon\}$. Moreover, assume the Lipschitz condition

$$\exists\, L > 0\, \forall\, s, s' \in \mathcal{S} \,:\, \Delta_\mathcal{R}(\varphi(s), \varphi(s')) \le L\, \Delta_\mathcal{S}(s, s') \tag{3.20}$$

to hold. Then, a finite memory $\mathcal{M}$ exists such that

$$\operatorname{diam}(\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s)) \overset{\text{df}}{=} \max\{\Delta_\mathcal{R}(r, r') \,|\, r, r' \in \widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s)\} \le 2\, L\, \varepsilon$$

for all $s \in \mathcal{S}$. □

**Proof.** Let $\varepsilon > 0$ and $\mathcal{S}' \subseteq \mathcal{S}$ satisfy (3.19) and define $\mathcal{M} = \bigcup_{s' \in \mathcal{S}'} \langle s', \varphi(s') \rangle$. For $s, s' \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s, s') = x \in D_\mathcal{S}$ we have $\Delta_\mathcal{S}(s, s') = f^{-1}(x)$. Thus, according to (3.20), $\sigma_\mathcal{R}(\varphi(s), \varphi(s')) \ge g(Lf^{-1}(x))$, which means $h_\Sigma(x) \ge g(Lf^{-1}(x))$ for all $x \in D_\mathcal{S}$. Now, consider some $s \in \mathcal{S}$. According to (3.19), the memory $\mathcal{M}$ contains a case $\langle s_0, r_0 \rangle$ such that $\Delta_\mathcal{S}(s, s_0) \le \varepsilon$. Hence, $h_\Sigma(\sigma_\mathcal{S}(s, s_0)) \ge g(Lf^{-1}(\sigma_\mathcal{S}(s, s_0))) \ge g(L\varepsilon)$, which means that $\Delta_\mathcal{R}(r_0, r') \le L\,\varepsilon$ for all $r' \in \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r_0)$. The result then follows from $\Delta_\mathcal{R}(r, r') \le \Delta_\mathcal{R}(r, r_0) + \Delta_\mathcal{R}(r_0, r')$ for all $r, r' \in \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r_0)$ and $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s) \subseteq \mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r_0)$. □

Since $\varphi(s) \in \widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$, Proposition 3.20 guarantees the existence of a case-based approximation of $\varphi$ which determines all outcomes up to a precision of $\delta = 2\, L\, \varepsilon$. The following corollaries follow immediately.

**Corollary 3.21.** Suppose the assumptions of Proposition 3.20 to hold true with "$\exists\, \varepsilon > 0$" in (3.19) replaced by "$\forall\, \varepsilon > 0$." Then, the mapping $\varphi$ can be approximated to any degree of accuracy $\delta > 0$ via (3.15) with a finite memory $\mathcal{M}$. □

**Corollary 3.22.** Let $\mathcal{S} \subseteq \mathfrak{Q}^p$ be bounded, $\mathcal{R} \subseteq \mathfrak{Q}^q$, and $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{R}}$ be defined by the corresponding Euclidean distances. Moreover, suppose that $\varphi$ satisfies (3.20) and that $\sigma_{\mathcal{S}} = f \circ \Delta_{\mathcal{S}}$ and $\sigma_{\mathcal{R}} = g \circ \Delta_{\mathcal{R}}$ with $f, g$ strictly decreasing. Then, $\varphi$ can be approximated to any degree of accuracy $\delta > 0$ via (3.15) with a finite memory $\mathcal{M}$. □

Assumption (3.19), which requires the existence of a finite cover of $\mathcal{S}$, cannot be dropped, as can easily be seen by constructing a counter-example with $\Delta_{\mathcal{S}}$ defined by $\Delta_{\mathcal{S}}(s, s') = 0$ for $s = s'$ and $\Delta_{\mathcal{S}}(s, s') = 1$ for $s \neq s'$ (and $\operatorname{card}(\mathcal{S}) = \aleph_0$). Likewise, (3.20) is necessary, as an example with $\varphi(s)$ defined on $[0, 1] \cap \mathfrak{Q}$ by $\varphi(s) = 1$ for $s = 0$ and $\varphi(s) = 0$ for $s > 0$ (and $\Delta_{\mathcal{R}}$ the standard metric) shows.

The discussion so far has shown that the inference scheme presented in Section 3.2 can basically be seen as a *set-valued* approximation method. The essential part of this inference procedure is realized in what we have called the *similarity space*, not in the instance space itself (cf. Fig. 3.2). That is, CBI is not directly based on the information provided at the system level. Rather, the concept of similarity, quantified in terms of similarity functions $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$, is exploited in order to transform this information into information which is represented at the similarity level. An approximation at the instance level is then derived within a two-stage process from inferences about the *similarity* of an unknown outcome to already observed ones.

It is this indirect derivation of approximations that constitutes the main difference between CBA and other approximation techniques. In fact, an implicit notion of similarity is also present in other methods, since the (local) transfer of observed outputs is generally based on the concept of *distance*. Typically, a (scalar) estimation of an unknown value $f(x)$ of a function $f$ is derived in the form of a weighted combination of training examples $f(x_1), \ldots, f(x_n)$, where the weight of an example $f(x_k)$ decreases with the distance of the associated point in the input space, $x_k$, to the query point $x$.[12] Consider an approximation of the form

$$\widehat{f}(x) = \frac{\sum_{k=1}^{n} K(x_k - x) \cdot f(x_k)}{\sum_{k=1}^{n} K(x_k - x)},$$

where $K(\cdot)$ is a *kernel function* (centered at 0), as an example.

In some approximation methods the observed outcomes $f(x_k)$ appear only implicitly, in the sense that they determine parameters of an approximating function. In a special version of locally weighted regression, for instance, the parameters of a linear function $\widehat{f}(\cdot)$ are determined such that

$$\sum_{k=1}^{n} (f(x_k) - \widehat{f}(x_k))^2 K(d(x, x_k))$$

[12] The input space must hence be endowed with a distance measure.

is minimized, where $d(\cdot)$ is a distance measure, and $K(\cdot)$ is a kernel function. The value of $f(\cdot)$ for the query point $x$ is then estimated by $\widehat{f}(x)$.

As a further example consider again the $k$-Nearest Neighbor ($k$NN) method (cf. Section 2.2) from which several instance-based learning algorithms have emerged. It derives predictions according to

$$\widehat{f}(x) = F(f(x_1), \ldots, f(x_k)),$$

where $f(x_1), \ldots, f(x_k)$ are the training examples associated with the $k$ points which are most similar to (or have the smallest distance from) the query point $x$. If $f(\cdot)$ is a numerical function, $F(\cdot)$ is often defined as a weighted average, i.e.

$$\widehat{f}(x) = \sum_{\jmath=1}^{k} \left( 1 - \frac{|x - x_\jmath|}{\sum_{\imath=1}^{k} |x - x_\imath|} \right) \cdot f(x_\jmath).$$

If $\mathrm{rg}(f)$ is discrete, $F(\cdot)$ generally returns the value which is most frequent among $f(x_1), \ldots, f(x_k)$.

As can be seen, the approximation methods outlined above are based on the same data as CBA, namely a set of observed values of a function (= outcomes) and some kind of similarity or distance relation between points (= inputs) in the input space. This data can be defined as an extension of the similarity structure (cf. Definition 3.13 and Fig. 3.3).

**Definition 3.23 (outcome structure).** Let $\Sigma$ be a CBI setup, $s_0$ a new input, and $\mathcal{M}$ the memory (2.29) associated with $\Sigma$. The set of values

$$\mathsf{OST}(\mathcal{M}, s_0) \overset{\mathrm{df}}{=} \mathsf{SST}(\mathcal{M}, s_0) \cup \{r_\jmath \,|\, 1 \leq \jmath \leq n\}$$

(together with $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$) defines the outcome structure of the CBI problem $\langle \Sigma, s_0 \rangle$. □

Usual approximation methods employ the outcome structure directly within one inference step. As opposed to this, the first step of the CBA scheme uses only the similarity structure, and the observed outcomes $r_k$ are called in for the second inference step.

The aforementioned difference becomes obvious, e.g., when comparing CBA to the $k$NN algorithm. Firstly, this algorithm applies the similarity measures directly at the instance level in order to find the most similar cases, whereas in CBA these measures are used for defining the similarity structure $h_\Sigma$. Secondly, the $k$NN method does also perform the inference step at the instance level, in the sense that predictions are derived directly from the observed outcomes. As opposed to this, CBA uses the given information for drawing inferences, not about outputs, but about similarities. It makes use of observed outcomes by more indirect means,

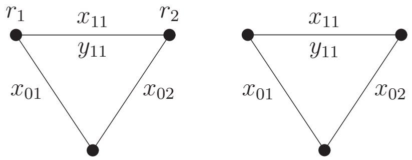


**Fig. 3.3.** The outcome (left) and similarity structure (right) of a CBI problem can be illustrated as a graph, where the nodes are associated with (information about) cases and the edges are labeled with information concerning the (similarity) relation between cases. This figure shows the graphs for a memory with two cases.

in the sense that each output defines an instantiation of a similarity constraint at the system level.

In connection with the $k$NN method it should also be observed that CBA (especially (3.16)) can be seen as an interesting *set-valued* version of this algorithm. As an advantage of CBA let us mention that it also takes the quality of the similarity structure into account when predicting an outcome. In fact, (3.15) will not be very constraining if this structure is poorly developed, thus indicating that the application of the NEAREST NEIGHBOR principle (and, hence, the original $k$NN method) does not seem advisable. We shall come back to this point in Chapter 4.

The following points deserve mentioning when comparing case-based to other local approximation methods. On the one hand, CBA is less demanding in the sense that it requires the specification of a similarity hypothesis, i.e., a relatively simple one-dimensional function, whereas other methods derive approximating functions with $\text{dom}(f) = \mathcal{S}$ and $\text{codom}(f) = \mathcal{R}$. Moreover, CBA still works if $\mathcal{S}$ and $\mathcal{R}$ are not as well-structured as certain number spaces, a situation regularly encountered within the context of CBR. In fact, the assignment of similarity degrees can then be seen as a reasonable quantification of the approximation problem. This kind of quantification will often be more obvious than a quantification of $\mathcal{S}$ and $\mathcal{R}$ which allows for deriving a good approximation $\widehat{f} : \mathcal{S} \longrightarrow \mathcal{R}$.

On the other hand, the transformation from a high-dimensional (instance) space into a low-dimensional (similarity) space is usually afflicted with a loss of information. This becomes especially apparent in connection with the (pseudo-)inverse of the similarity measure $\sigma_{\mathcal{R}}$. In fact, this transformation will generally be a *set-valued* mapping.

In any case, a comparison between (indirect) case-based and direct approximation methods remains a difficult (if not meaningless) task. Firstly, the success of any approximation method largely depends on the application and properties of the data.[13] Thus, it will generally not be possible to qualify one approach as being superior in comparison to other methods. Secondly, a case-based approximation is not scalar-valued but derives set-valued approximations which either cover the unknown outcome or, as will be seen in Section 3.4, define some kind of confidence region. Thus, the usefulness of an approximation method will also depend on whether the problem at hand requires an estimation in the form of a scalar value $\widehat{r}_0$ or whether it is important to have information about $r_0$ in the form of outer bounds. As can be seen, the aforementioned differences between case-based and direct approximation suggest to combine (rather than to compare) these approaches.

### 3.3.2 Local similarity profiles

In Section 3.2.1, it has already been pointed out that CBA is *local* in the sense that the information provided by different cases is processed and combined independently.[14] It is, however, *global* in the sense that the similarity profile represents information which holds true for the complete similarity space. In fact, the constraint $\mathcal{N}_{h_\Sigma(\sigma_\mathcal{S}(s,s_0))}(r)$ provided by a case $\langle s, r\rangle$ for the prediction of an unknown outcome $\varphi(s_0)$ contains a local component, namely the case $\langle s, r\rangle$ itself, as well as a global component, namely the similarity hypothesis $h$. CBA can thus be characterized as a local processing of global information.

Often, the CBI assumption is not satisfied equally well for all parts of the instance space $\mathcal{S}\times\mathcal{R}$.[15] The global validity of the similarity profile might then prevent one from defining tight bounds for those regions where the CBI hypothesis actually applies rather well. In fact, a globally admissible similarity hypothesis might lead to (local) predictions which are unnecessarily imprecise. This is illustrated by the following simple example.

Example 3.24. Let $\mathcal{S} = \mathcal{R} = [-1, 1] \setminus \{0\}$,[16] $\varphi(s) = -1$ if $-1 \leq s < 0$, and $\varphi(s) = 1$ if $0 < s \leq 1$. Moreover, let $\sigma_\mathcal{S}(u, v) = \sigma_\mathcal{R}(u, v) = 1 - |u - v|/2$. Obviously, for all $1 \neq x \in D_\mathcal{S}$ there are $s, s' \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s, s') = x$ and $\sigma_\mathcal{R}(\varphi(s), \varphi(s')) = 0$. We hence have $h_\Sigma(x) = 0$ for all $x \in D_\mathcal{S} \setminus \{1\}$, which means that $\widehat{\varphi}_{h_\Sigma,\mathcal{M}}(s_0) = [-1, 1]$ if $\langle s_0, \varphi(s_0)\rangle \notin \mathcal{M}$. □

Loosely speaking, a CBI strategy is not applicable in Example 3.24 because the CBI hypothesis is not globally valid. Still, it seems desirable to make use of the observation that this assumption is satisfied at least *locally*. One possibility of doing this is to partition the set $\mathcal{S}$ of inputs and to derive respective local

[13] Recall the selective superiority problem mentioned in footnote 9.

[14] CBA is also local in the sense that it is a local approximation method. These two meanings of locality should not be confused.

[15] In a game playing context, for instance, the CBI principle hardly applies to certain "tactical" situations [310].

[16] More specifically, to comply with our formal framework, we should set $\mathcal{S} = \mathcal{R} = ([-1, 1] \cap \mathfrak{Q}) \setminus \{0\}$.

approximations.[17] In Example 3.24, it suggests itself to partition $\mathcal{S}$ into $[-1,0)$ and $(0,1]$. However, since $\varphi$ is generally unknown, the definition of such a partition will not always be obvious, all the more if $\mathcal{S}$ is non-numerical. Here, we consider a second possibility, namely that of associating an individual similarity profile with each input of the memory. This approach is somehow comparable to the use of local kernels in kernel-based density estimation [385], and to the use of *local metrics* in $k$NN algorithms and instance-based learning (e.g., metrics which allow feature weights to vary as a function of the instance [342, 157, 9, 311]). It leads us to introduce the concept of a *local similarity profile*.

**Definition 3.25 (local similarity profile).** Consider a CBI setup $\Sigma$ and let $s \in \mathcal{S}$. We define $h^s_\Sigma : D_\mathcal{S} \longrightarrow [0,1]$ by the mapping

$$x \mapsto \inf_{s' \in \mathcal{S},\, \sigma_\mathcal{S}(s,s')=x} \sigma_\mathcal{R}(\varphi(s), \varphi(s')).$$

This function is called the local similarity profile associated with $s$, or the $s$-similarity profile of $\Sigma$. A collection $h^\mathcal{M}_\Sigma = \{h^s_\Sigma \,|\, s \in \mathcal{M}^\downarrow\}$ of local profiles is referred to as the local $\mathcal{M}$-similarity profile. □

The following relations hold between the different types of similarity profiles:

$$h_\Sigma = \bigwedge_{s \in \mathcal{S}} h^s_\Sigma, \quad h^\mathcal{M}_\Sigma = \bigwedge_{s \in \mathcal{M}^\downarrow} h^s_\Sigma.$$

That is, the similarity profile $h_\Sigma$ and $\mathcal{M}$-similarity profile $h^\mathcal{M}_\Sigma$ are lower envelopes of the class of local profiles associated with inputs in $\mathcal{S}$ and $\mathcal{M}^\downarrow$, respectively. Consequently, $h_\Sigma \leq h^\mathcal{M}_\Sigma \leq h^s_\Sigma$ for all memories $\mathcal{M}$ and inputs $s \in \mathcal{M}^\downarrow$.

As can be seen, a local similarity profile is closely related to the idea of an $\mathcal{M}$-similarity profile. In fact, an $s$-profile corresponds to the $\mathcal{M}$-profile with $\mathcal{M}^\downarrow = (s)$. Besides, a class of local profiles will generally be specified – by means of respective learning methods (cf. Section 3.4) – for a memory which does not change frequently. In connection with approximation methods, the inputs which constitute the memory and for which local profiles are defined play a role somewhat similar to the so-called *knots* in, say, approximation with spline functions, and the local profiles correspond to basis functions.

Given a hypothesis $h^\mathcal{M} = \{h^s \,|\, s \in \mathcal{M}^\downarrow\}$ related to a local $\mathcal{M}$-similarity profile and a new input $s_0 \in \mathcal{S}$, the inference scheme (3.2) is replaced by

$$\varphi(s_0) \in \widehat{\varphi}_{h^\mathcal{M},\mathcal{M}}(s_0) \overset{\text{df}}{=} \bigcap_{\langle s,r \rangle \in \mathcal{M}} \mathcal{N}_{h^s(\sigma_\mathcal{S}(s,s_0))}(r). \tag{3.21}$$

The respective case-based approximation, i.e., the local counterpart to (3.15), is called a *local case-based approximation*:

[17] This idea is related to that of *feature space partitioning* in classification [77]. See also [261] for a related idea in connection with memory-based learning.

$$\widehat{\varphi}_{h^{\mathcal{M}},\mathcal{M}} : s \mapsto \bigcap_{\langle s',r'\rangle \in \mathcal{M}} \mathcal{N}_{h^{s'}(\sigma_{\mathcal{S}}(s,s'))}(r').$$

Example 3.26. Consider again Example 3.24 and suppose that the memory $\mathcal{M}$ contains the cases $\langle -1,-1\rangle$ and $\langle 1,1\rangle$. The respective local profiles are given by

$$x \mapsto \begin{cases} 1 & \text{if} \;\; 1/2 \leq x \leq 1 \\ 0 & \text{if} \;\; 0 \leq x < 1/2 \end{cases} .$$

These two profiles can already guarantee an exact representation of $\varphi$. That is, $\widehat{\varphi}_{h^{\mathcal{M}}_{\Sigma}}(s) = \{\varphi(s)\}$ for all $s \in \mathcal{S}$ with $\mathcal{M} = (\langle -1,-1\rangle, \langle 1,1\rangle)$. □

Note that a local profile indicates the validity of the CBI hypothesis for *individual* cases. That is, the local profile associated with an input $s \in \mathcal{S}$ can be utilized for rating the *quality* of the case $\langle s, \varphi(s)\rangle$.[18] An input with a strongly developed local profile (i.e., its outcome is locally representative) will generally support precise predictions, whereas an input with a poorly developed profile will hardly be useful from the viewpoint of CBA. Local profiles might hence serve as a (complementary) criterion for selecting "competent" cases to be stored in (or removed from) the memory [357]. It should be noted, however, that the similarity profile can only be taken as an indication of the precision of predictions. In fact, the predictions also depends on the neighborhood structure of $\mathcal{R}$. For instance, it is quite possible that $\text{card}(\mathcal{N}_{\alpha}(r)) < \text{card}(\mathcal{N}_{\beta}(r'))$ for two outcomes $r \neq r'$, even though $\beta < \alpha$.

## 3.4 Learning similarity hypotheses

### 3.4.1 The learning task

The inference scheme (3.2) reveals that CBI can essentially be seen as an *instance-based* approach. Still, it also contains a *model-based* component, namely the similarity hypothesis $h$. Consequently, *learning* can be realized in (at least) two ways in CBI: By storing new cases in the memory and by estimating the similarity profile. Here, we concentrate on the latter (model-based) aspect.

**Definition 3.27** (CBL). Consider a CBI setup $\Sigma$ with a memory

$$\mathcal{M} \subseteq \mathcal{D} = \mathcal{D}_N = (c_1, \ldots, c_N),$$

where $\mathcal{D}$ denotes the sequence of cases which have been encountered so far (these are the first $N$ cases, given the assumption that cases arrive successively). Moreover, let $\mathcal{H}$ be a *hypothesis space* of functions $h : [0,1] \longrightarrow [0,1]$. The task of *case-based learning* (CBL) is understood as deriving an optimal hypothesis $h_* \in \mathcal{H}$ from the data given. □

[18] See Section 4.6 for a more detailed discussion of the assessment of cases.

Observe that different similarity measures define different similarity structures of the system under consideration and that the measures originally chosen might not be optimal in the sense that similarity structures induced by alternative measures are, in a certain sense, more suitable for CBI. Suppose, for instance, that we have measures $(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$ and $(\sigma'_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ and let $\widehat{\varphi}_{h,\mathcal{M}}$ resp. $\widehat{\varphi}'_{h,\mathcal{M}}$ denote the case-based approximations induced by these measures via (3.15) with $h = h_{\Sigma}$. If $\widehat{\varphi}_{h,\mathcal{M}}(s) \subseteq \widehat{\varphi}'_{h,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$, then $(\sigma'_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ should not (at least not strictly) be preferred to $(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. This gives rise to defining a partial order relation on a class of measures. Therefore, it might also be reasonable to allow for the adaptation of similarity measures. The problem of CBL can thus be extended as follows.

**Definition 3.28 (extended CBL problem).** Let a set $\mathcal{S}$ of inputs, a set $\mathcal{R}$ of outputs, and a memory $\mathcal{M} \subseteq \mathcal{D} = (c_1, \ldots, c_N)$ be given, where $\mathcal{D}$ denotes the sequence of cases which have been encountered so far. Moreover, let $\mathcal{H}$ be a class of functions $h : [0,1] \longrightarrow [0,1]$ and $\mathcal{H}_{\mathcal{S}}$, $\mathcal{H}_{\mathcal{R}}$ classes of similarity measures over $\mathcal{S}$ and $\mathcal{R}$, respectively. The task of (extended) CBL is defined as searching the hypothesis space $\mathcal{H} \times \mathcal{H}_{\mathcal{S}} \times \mathcal{H}_{\mathcal{R}}$ for an optimal hypothesis $h_* = (h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. □

Remark 3.29. Relating the interpretation of a similarity hypothesis $h$ (resp. a similarity profile $h_{\Sigma}$) to the idea of modifying the measure $\sigma_{\mathcal{S}}$ has already been suggested in Remark 3.4. If $h$ is strict, such a modification corresponds to a "stretching" and "squeezing" of the similarity scale underlying $\sigma_{\mathcal{S}}$. Moreover, the modification is restricted in the sense that the original measure $\sigma_{\mathcal{S}}$ and its modified version $\sigma'_{\mathcal{S}}$ are coherent in the sense of (3.17). As opposed to this, a non-monotone hypothesis additionally puts the similarity degrees $x \in D_{\mathcal{S}}$ in a different order, which corresponds to a re-arranging of the (ordinal) similarity scale $D_{\mathcal{S}}$. Then, (3.17) holds true only with $\leq$ replaced by the equality relation. In other words, two inputs $s_1, s_2$ which are more similar than the inputs $s_3, s_4$ according to $\sigma_{\mathcal{S}}$ might be seen as being less similar according to $\sigma'_{\mathcal{S}}$. Now, one possibility to approach the extended CBL problem is to allow for a re-arranging of the similarity scale underlying $\sigma_{\mathcal{R}}$ as well, i.e., to allow for replacing $\sigma_{\mathcal{R}}$ by $\sigma'_{\mathcal{R}} = m \circ \sigma_{\mathcal{R}}$ for some $m : [0,1] \longrightarrow [0,1]$. A similarity hypothesis $h$ is then related to $(\sigma_{\mathcal{S}}, \sigma'_{\mathcal{R}})$ instead of $(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$. In connection with the extended CBL problem, this amounts to defining $\mathcal{H}_{\mathcal{R}}$ as the class of all measures which can be written in the form $m \circ \sigma_{\mathcal{R}}$. □

Definition 3.27 has not commented on the criteria which decide on the optimality of hypotheses. In order to derive such criteria we fall back on two principles. The first one is the obvious demand that an optimal hypothesis $h_*$ should be consistent with observed data in the sense that (3.1) is satisfied at least for elements of $\mathcal{D}$, i.e.

$$(\sigma_{\mathcal{S}}(s, s') = x) \ \Rightarrow \ (\sigma_{\mathcal{R}}(r, r') \geq h_*(x)) \tag{3.22}$$

should hold true for all $\langle s, r\rangle, \langle s', r'\rangle \in \mathcal{D}$. This *consistency principle* is closely related to the *inductive learning hypothesis* in machine learning. Namely, we suspect a hypothesis $h$, which is consistent with a large number of observations, also to be consistent with the overall similarity structure of the system (in the sense that it is admissible). Observe that (3.22) implies $\varphi(s) \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all $s$ with $\langle s, \varphi(s) \rangle \in \mathcal{D}$ and $\mathcal{M} \subseteq \mathcal{D}$. Again, we may assume that a mapping which defines an outer approximation of $\varphi|\mathcal{S}'$ for a (large) subset $\mathcal{S}' \subseteq \mathcal{S}$ also defines an outer approximation of the complete mapping $\varphi = \varphi|\mathcal{S}$. We denote by $\mathcal{H}_{\mathcal{D}} \subseteq \mathcal{H}$ the class of hypotheses which are consistent with a set $\mathcal{D}$ of cases in the sense of (3.22).

As will be seen in Section 3.4.3, it may become necessary to weaken the aforementioned consistency principle. In fact, testing consistency of a hypothesis according to (3.22) requires the consideration of all pairs $(c, c') \in \mathcal{D} \times \mathcal{D}$ of cases. However, as suggested by Definition 3.27, the memory $\mathcal{M}$ of stored cases will generally be a (proper) subset of the set $\mathcal{D}$ of *successively* encountered cases. It is hence not possible to take the tuple, say, $(c_1, c_N)$ into consideration if $c_1$ was not stored long enough and has been removed before the arrival of $c_N$. Thus, a weaker version of the consistency principle should require (3.22) to hold true for all

$$(c, c') \in \mathcal{C} = \mathcal{C}_N \overset{\mathrm{df}}{=} \bigcup_{1 \leq n \leq N-1} \mathcal{M}_n \times (c_{n+1}),$$

where $\mathcal{M}_n$ denotes the memory after the observation of the $n$-th case $c_n$. We denote by $\mathcal{H}_{\mathcal{C}}$ the class of hypotheses which are consistent with $\mathcal{D}$ in this weaker sense. Thus, we generally have $\mathcal{H}_{\mathcal{D}} \subseteq \mathcal{H}_{\mathcal{C}} \subseteq \mathcal{H}_{\mathcal{M}}$, where $\mathcal{H}_{\mathcal{M}}$ is defined in a canonical way.

In order to motivate the second principle recall that the case-based approximation (3.15), which is induced by a hypothesis $(h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$ and a memory $\mathcal{M}$, can be seen as a simplified representation of the system structure $\varphi$. Indeed, $\widehat{\varphi}_{h,\mathcal{M}}$ is represented by card($\mathcal{M}$) cases and the hypothesis $(h, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}})$, whereas the representation of $\varphi$ – if it cannot be expressed in closed form – requires the enumeration of the complete set

$$\mathcal{D}^* \overset{\mathrm{df}}{=} \{\langle s, \varphi(s) \rangle \,|\, s \in \mathcal{S}\}$$

of cases. Of course, in passing from $\varphi$ to $\widehat{\varphi}_{h,\mathcal{M}}$ it is usually unavoidable to loose some information. The corresponding increase in uncertainty is reflected by the fact that $\widehat{\varphi}_{h,\mathcal{M}}$ is a *set-valued* mapping and that we will generally have $\{\varphi(s)\} \subsetneq \widehat{\varphi}_{h,\mathcal{M}}(s)$ for at least some inputs $s \in \mathcal{S}$. According to the *principle of minimum uncertainty*, which is one of the general principles of systems theory, one should, among a set of candidates, accept only those simplifications of a system for which the increase in uncertainty is minimal [231]. Thus, let $U$ be some measure which quantifies the uncertainty associated with $\widehat{\varphi}_{h,\mathcal{M}}$.[19] A hypothesis $h_*$ is then *optimal* if $h_* \in \mathcal{H}_{\mathcal{C}}$ and $U\left(\widehat{\varphi}_{h_*,\mathcal{M}}\right) \leq U\left(\widehat{\varphi}_{h,\mathcal{M}}\right)$ holds true for all $h \in \mathcal{H}_{\mathcal{C}}$. We denote by $\mathcal{H}_* \subseteq \mathcal{H}_{\mathcal{C}}$ the class of all optimal hypotheses. Of course, this definition does neither guarantee the existence nor the uniqueness of an optimal hypothesis.

[19] Various proposals for such uncertainty measures can be found in systems science literature.

In connection with the learning of hypotheses it makes sense to consider *admissibility* as a further property which is more restricting than consistency. We denote by $\mathcal{H}^*$ the class of optimal admissible hypotheses. Thus, $\mathcal{H}^*$ consists of those uncertainty minimizing hypotheses $h^*$ which are consistent with $\mathcal{D}^*$.[20]

Let us now consider the CBL problem in its basic form. Of course, deriving the uncertainty $U(\widehat{\varphi}_{h,\mathcal{M}})$ associated with a hypothesis $h$ is intractable if it requires the computation of the complete mapping $\widehat{\varphi}_{h,\mathcal{M}}$. Observe, however, that any reasonable measure $U$ should satisfy $U(\widehat{\varphi}_{h,\mathcal{M}}) \leq U(\widehat{\varphi}_{h',\mathcal{M}})$ if $\widehat{\varphi}_{h,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h',\mathcal{M}}(s)$ for all $s \in \mathcal{S}$. Since the latter holds true if $h' \leq h$, $U$ should be consistent with the partial order defined by $\leq$ over $\mathcal{H}$.

**Observation 3.30.** Suppose the hypothesis space $\mathcal{H}$ to satisfy $h \equiv 0 \in \mathcal{H}$ and $(h, h' \in \mathcal{H}) \Rightarrow (h \vee h' \in \mathcal{H})$, where $h \vee h'$ is defined by the mapping $x \mapsto \max\{h(x), h'(x)\}$. Moreover, suppose the measure $U$ to satisfy

$$(h' \leq h) \;\Rightarrow\; (U(\widehat{\varphi}_{h,\mathcal{M}}) \leq U(\widehat{\varphi}_{h',\mathcal{M}}))$$

for all $h, h' \in \mathcal{H}$ and memories $\mathcal{M}$. Then, a unique optimal hypothesis $h_* \in \mathcal{H}$ exists, and $\mathcal{H}_{\mathcal{C}} = \{h \in \mathcal{H} \,|\, h \leq h_*\}$. □

Given the assumptions of Observation 3.30, CBL can be realized as a *candidate-elimination* algorithm [269], where $h_*$ is a compact representation of the *version space*, i.e., the subset $\mathcal{H}_{\mathcal{C}}$ of hypotheses from $\mathcal{H}$ which are consistent with the training examples.

Note that (3.22) guarantees consistency in the "empirical" sense that $r \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all observed cases $\langle s, r\rangle \in \mathcal{D}$. Still, one might think of demanding furthermore a kind of "logical" consistency, namely $\widehat{\varphi}_{h,\mathcal{M}}(s') \neq \emptyset$ for the set of all possible inputs $s' \in \mathcal{S}$. Of course, this additional demand would greatly increase the complexity of testing consistency. Moreover, the assumptions of Observation 3.30 would no longer guarantee the existence of a unique optimal hypothesis.

Since two hypotheses $h$ and $h'$ are only comparable for the same underlying similarity measures (cf. Remark 3.7), the above remarks do not apply to the extended CBL problem. Thus, considering the maps $\widehat{\varphi}_{h,\mathcal{M}}$ themselves cannot be avoided in this case. Nevertheless, one can think of efficient (heuristic) approaches for realizing corresponding learning procedures. A value $U(\widehat{\varphi}_{h,\mathcal{M}})$ might be approximated, for instance, by some value $\widehat{U}(\{\widehat{\varphi}_{h,\mathcal{M}}(s) \,|\, s \in \mathcal{S}'\})$ derived from a sample $\mathcal{S}' \subseteq \mathcal{S}$. The usefulness of different (generalized) learning procedures will, however, highly depend on characteristics of the similarity measures and the way in which these measures can be adapted, i.e., on the classes $\mathcal{H}_{\mathcal{S}}$ and $\mathcal{H}_{\mathcal{R}}$. In this section, we shall restrict ourselves to the basic version of the CBL problem.

[20] Observe that $\mathcal{H}^* \subseteq \mathcal{H}_*$ does generally not hold.

### 3.4.2 A learning algorithm

Let hypotheses be represented by step functions

$$h : x \mapsto \sum_{k=1}^{m} \beta_k \cdot \mathbb{I}_{A_k}(x), \tag{3.23}$$

where $A_k = [\alpha_{k-1}, \alpha_k)$ for $1 \leq k \leq m-1$, $A_m = [\alpha_{m-1}, \alpha_m]$ and $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$ defines a partition of $[0,1]$.[21] The hypothesis $h$ can then be associated with a set of rules (implications) of the form

$$(\sigma_{\mathcal{S}}(s,s') \in A_k) \;\Rightarrow\; (\sigma_{\mathcal{R}}(\varphi(s), \varphi(s')) \geq \beta_k). \tag{3.24}$$

Observe that by simply defining one interval for each element $x \in D_{\mathcal{S}}$, $h_{\Sigma}$ itself can be seen as a step function if $\mathcal{S}$ is finite. A combination (3.24) of such similarity degrees seems still reasonable if $\mathcal{S}$ is not finite (or even if $\operatorname{card}(\mathcal{S})$ is large).

The class $\mathcal{H}_{step}$ of functions (3.23), defined for a fixed partition, does obviously satisfy the assumptions of Observation 3.30. The optimal hypothesis $h_*$ is defined by the values

$$\beta_k \overset{\text{df}}{=} \min_{(s,s') \in \mathcal{C}^{\downarrow}, \sigma_{\mathcal{S}}(s,s') \in A_k} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s')) \tag{3.25}$$

for $1 \leq k \leq m$, where $\min \emptyset \overset{\text{df}}{=} 1$ by convention; see Fig. 3.4 for an illustration. Since this hypothesis is directly derived from the case base $\mathcal{M}$, we also call it the *empirical similarity profile*.

Now, suppose that $\mathcal{M}$ is the current memory and that a new case $c_0 = \langle s_0, r_0 \rangle$ has been observed. Updating $h_*$ can then be accomplished by passing the iteration

$$\beta_{\kappa(s_0,s_\jmath)} = \min\{\beta_{\kappa(s_0,s_\jmath)}, \sigma_{\mathcal{R}}(r_0, r_\jmath)\} \tag{3.26}$$

for $1 \leq \jmath \leq \operatorname{card}(\mathcal{M})$; the index $1 \leq \kappa(s,s') \leq m$ is defined for inputs $s, s' \in \mathcal{S}$ by $\kappa(s,s') = k \overset{\text{df}}{\Leftrightarrow} \sigma_{\mathcal{S}}(s,s') \in A_k$. As (3.26) shows, the representation (3.23) is computationally efficient. In fact, the time complexity of updating a hypothesis is linear in the size of the memory.[22] In other words, the model-based part of learning in CBI is not critical from a computational point of view. We refer to the algorithm defined by (3.26) as CBLA and denote by $\text{CBLA}(\mathcal{C})$ the hypothesis (3.25).

For obvious reason we call $h^* \in \mathcal{H}_{step}$ defined by

$$\beta_k^* \overset{\text{df}}{=} \inf_{x \in D_{\mathcal{S}} \cap A_k} h_{\Sigma}(x) \tag{3.27}$$

$(1 \leq k \leq m)$ the *optimal admissible* hypothesis. Since admissibility (in the sense of Definition 3.2) implies consistency, we have $h^* \leq h_*$.

[21] In Section 3.3.1 we have hinted at the *ordinal* character of the similarity measures $\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$. In connection with the representation of hypotheses according to (3.23) it should, therefore, be noticed that a scaling of $\sigma_{\mathcal{S}}$ might influence the optimal similarity hypothesis if the underlying partition is assumed to be *fixed*.

[22] We assume that $\kappa$ is computed in constant time.

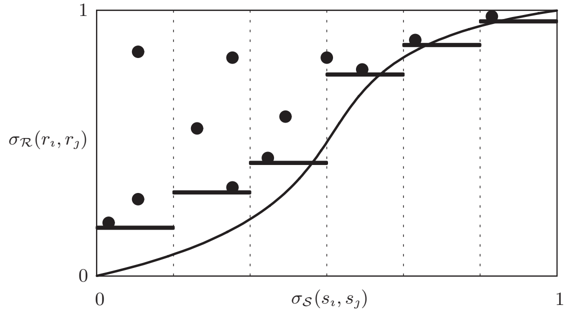


**Fig. 3.4.** Each pair of observed cases $\langle s_\imath, r_\imath\rangle$ and $\langle s_\jmath, r_\jmath\rangle$ contributes a point $(x, y)$ in the "similarity space", where $y = \sigma_{\mathcal{S}}(s_\imath, s_\jmath)$ and $r = \sigma_{\mathcal{R}}(r_\imath, r_\jmath)$. By definition, these points are located above the similarity profile, which is here shown by the solid curve. The optimal similarity hypothesis $h_*$ is given by the step function indicated by the solid horizontal lines.

Remark 3.31. Assuming the CBI hypothesis to hold true in the strict sense restricts the class $\mathcal{H}_{step}$ to the class $\mathcal{H}^{\uparrow}_{step}$ of non-decreasing step functions, which is also closed under $\vee$. Consider a hypothesis $h_* \in \mathcal{H}_{step}$ represented by values $(\beta_1, \ldots, \beta_m)$. Moreover, denote by $h^{\uparrow}_* \in \mathcal{H}^{\uparrow}_{step}$ the corresponding strict hypothesis represented by values $(\beta^{\uparrow}_1, \ldots, \beta^{\uparrow}_m)$. The relation between $h_*$ and $h^{\uparrow}_*$ is obviously given by $\beta^{\uparrow}_k = \min\{\beta_\jmath \,|\, k \leq \jmath \leq m\}$ for all $1 \leq k \leq m$. Thus, an optimal strict hypothesis can always be derived easily from $h_*$. □

Remark 3.32. If a similarity hypothesis $h$ is defined by a step function, the same is actually true for a case-based approximation $\widehat{\varphi}_{h,\mathcal{M}}$ itself. Namely, for $s, s' \in \mathcal{S}$ we have $\widehat{\varphi}_{h,\mathcal{M}}(s) = \widehat{\varphi}_{h,\mathcal{M}}(s')$ if $\kappa(s, s_\imath) = \kappa(s', s_\imath)$ for all $1 \leq \imath \leq n$. A corresponding equivalence relation on $\mathcal{S} \times \mathcal{S}$, where each equivalence class is identified by some vector $(k_1, \ldots, k_n)$ of indices $k_\jmath = \kappa(s, s_\jmath) \in \{1, \ldots, m\}$, offers some interesting possibilities of representing the mapping $\widehat{\varphi}_{h,\mathcal{M}}$ and deriving values thereof. For instance, since $\widehat{\varphi}_{h,\mathcal{M}}(s) = \widehat{\varphi}_{h,\mathcal{M}}(s')$ whenever $s$ and $s'$ are elements of the same equivalence relation, the values associated with the equivalence classes might be computed in advance and stored by means of an adequate data structure. The derivation of a value $\widehat{\varphi}_{h,\mathcal{M}}(s)$ then reduces to a (simple) "look-up" procedure. Admittedly, the number $m^n$ of (potential) classes is generally extremely large, even though most of them will be empty. □

### 3.4.3 Properties of case-based learning

We shall now consider an iterative scheme which is in accordance with the idea of CBI as a repeated process of problem solving and learning. This case-based learning process, called CBLP and outlined in Algorithm 1, is based on a random sequence $(S_N)_{N\geq 1}$ of inputs $S_N \in \mathcal{S}$ which are independent and identically distributed according to $\mu_{\mathcal{S}}$, and a sequence $p = (p_N)_{N\geq 1} \in [0,1]^\infty$.

**Algorithm 1** CBLP

```
Input: a sequence of query inputs
Output: a sequence of estimation for outputs
$\mathcal{M}_0 = \emptyset$, $h_0 \equiv 1$
$N = 0$
repeat
    compute $\widehat{r}_{N+1} = \widehat{\varphi}_{h_N,\mathcal{M}_N}(s_{N+1})$
    solve-problem($s_{N+1}, \widehat{r}_{N+1}$)
    $h_{N+1} =$ update($h_N, c_{N+1}, \mathcal{M}_N$)
    $\mathcal{M}_{N+1} = \begin{cases} \mathcal{M}_N \cup (c_{N+1}) & \text{with probability } p_{N+1} \\ \mathcal{M}_N & \text{with probability } 1 - p_{N+1} \end{cases}$
    $N = N + 1$
until no more queries exist
```

Here, `solve-problem` is a procedure in which the prediction $\widehat{r}_{N+1}$ is used for supporting the derivation of the true outcome $\varphi(s_{N+1})$. Moreover, the procedure `update`$(h_N, c_{N+1}, \mathcal{M}_N)$ returns the hypothesis obtained from $h_N$ by passing the iteration (3.26) for $\mathcal{M}_N$ and the case $c_{N+1} = \langle s_{N+1}, \varphi(s_{N+1})\rangle$. Observe that CBLP guarantees $h_N = \text{CBLA}(\mathcal{C}_N)$ but that we generally have $h_N \neq \text{CBLA}(\mathcal{D}_N \times \mathcal{D}_N)$. The probabilistic extension of the memory in CBLP takes into account that adding all observations to $\mathcal{M}$, i.e., taking $p \equiv 1$, might not be advisable [353]. Of course, efficient problem solving will generally assume a more sophisticated strategy for the instance-based aspect of learning, i.e., for maintaining the memory of cases. It might be reasonable, e.g., to take the "quality" of individual cases into account and to allow for removing already stored cases from the memory [355, 286]. Nevertheless, the probabilistic extension in CBLP allows for gaining insight into theoretical properties of the learning scheme. Observe that $p_N = 0$ for $N \geq N_0$ (with $N_0$ being a constant number) comes down to using a fixed memory $\mathcal{M}$.

Given a CBI setup and the sequence $(p_N)_{N\geq 1}$, the hypotheses $h_N$ induced by CBLP are random functions with well-defined (even though tremendously complicated) distributions. We are now going to derive some important properties of the sequence $(h_N)_{N\geq 1}$. It goes without saying that one of the first questions arising in connection with our learning scheme concerns the relation between $(h_N)_{N\geq 1}$ and the optimal admissible hypothesis $h^*$.

**Proposition 3.33.** Suppose $p \geq \delta > 0$, i.e., $p_N \geq \delta$ for all $N \in \mathfrak{N}$, and let $(h_N)_{N\geq 1}$ be the sequence of hypotheses induced by CBLP. Then, $h_N \searrow h^*$ stochastically as $N \to \infty$. That is, $h_N \geq h^*$ for all $N \in \mathfrak{N}$ and

$$\mathbb{P}(\|h_N - h^*\|_\infty \geq \varepsilon) \to 0$$

for all $\varepsilon > 0$. □

**Proof.** From the definition of $h^*$ and the updating scheme (3.26) it becomes obvious that $h^* \leq h_N$ for all $N \geq 1$ and that the sequence of functions $(h_N)_{N \geq 0}$ is decreasing. Let $\varepsilon > 0$ and consider some $1 \leq k \leq m$. According to (3.27), there is some $x \in A_k$ such that $|h_\Sigma(x) - \beta^*_k| < \varepsilon/2$. Since we have $h_\Sigma(x) = \inf\left\{\sigma_\mathcal{R}(\varphi(s), \varphi(s')) \,|\, s, s' \in \mathcal{S}, \sigma_\mathcal{S}(s, s') = x\right\}$, there are also values $s_{k_1}, s_{k_2} \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s_{k_1}, s_{k_2}) = x$ and $|\sigma_\mathcal{R}(\varphi(s_{k_1}), \varphi(s_{k_2})) - h_\Sigma(x)| < \varepsilon/2$. Hence, $|\sigma_\mathcal{R}(\varphi(s_{k_1}), \varphi(s_{k_2})) - \beta^*_k| < \varepsilon$. This implies $|h_{\mathcal{M}_N}(x) - \beta^*_k| < \varepsilon$ as soon as the memory $\mathcal{M}_N$ contains the inputs $s_{k_1}$ and $s_{k_2}$, where $h_{\mathcal{M}_N} = \text{CBLA}(\mathcal{M}_N)$. Since this argumentation applies to all $1 \leq k \leq m$ and since $h^* \leq h_N \leq h_{\mathcal{M}_N}$, we obtain

$$\|h_N - h^*\|_\infty \leq \|h_{\mathcal{M}_N} - h^*\|_\infty = \max_{0 \leq x \leq 1} |h_{\mathcal{M}_N}(x) - h^*(x)| < \varepsilon$$

if $\mathcal{M}_N$ contains the (at most $2\,m$) inputs $s_{k_1}, s_{k_2}$ ($1 \leq k \leq m$). Since $\mu_\mathcal{S}(s_{k_1}) > 0$ and $\mu_\mathcal{S}(s_{k_2}) > 0$ for all $1 \leq k \leq m$ and $p_N \geq \delta > 0$ for all $N \in \mathfrak{N}$, the probability for this tends toward 1 for $N \to \infty$. □

Observe that the stochastic convergence (from above) of the hypotheses $(h_N)_{N \geq 0}$ toward $h^* \in \mathcal{H}_{step}$, which is guaranteed by Proposition 3.33, does not imply that $h_N(x) \to h_\Sigma(x)$ for all $x \in D_\mathcal{S}$. In fact, it might happen that $h^*|D_\mathcal{S}$ is already a poor approximation of $h_\Sigma$ (at least in the strong sense of the $\|\cdot\|_\infty$ metric) regardless of the (finite) partition underlying the definition of the hypothesis space $\mathcal{H}_{step}$. The following example shows that this cannot be avoided even if the system $(\mathcal{S}, \mathcal{R}, \varphi)$ satisfies strong structural assumptions:

Example 3.34. Let $\mathcal{S} = \{s_k = k - (1/2)^k \,|\, k \in \mathfrak{N}_0\}$, $\mathcal{R} = \{0, 1\}$, and

$$\varphi(s_k) = \begin{cases} 0 & \text{if} \quad \lfloor k/2 \rfloor \text{ is odd} \\ 1 & \text{if} \quad \lfloor k/2 \rfloor \text{ is even} \end{cases} .$$

Moreover, let $\sigma_\mathcal{S}(s, s') = |s - s'|^{-1}$ and $\sigma_\mathcal{R}(r, r') = 1 - |r - r'|$ (and note that $\varphi : (\mathcal{S}, |\cdot|) \longrightarrow (\mathcal{R}, |\cdot|)$ does even satisfy a Lipschitz condition). Now, for $\alpha_k = 2^k/(2^k + 1)$ ($k \in \mathfrak{N}$) there are exactly two inputs $s, s' \in \mathcal{S}$ such that $\sigma_\mathcal{S}(s, s') = \alpha_k$, namely $s = s_{k-1}$ and $s' = s_k$ (or vice versa). Thus, we have

$$h_\Sigma(\alpha_k) = \sigma_\mathcal{R}(\varphi(s_{k-1}), \varphi(s_k)) = \begin{cases} 1 & \text{if} \quad \text{k is odd} \\ 0 & \text{if} \quad \text{k is even} \end{cases} .$$

Obviously, each finite partition of $[0, 1]$ contains an interval $A$ such that $\alpha_k, \alpha_{k+1} \in A$ for some $k \geq 1$. Consequently, $h^*|A \equiv 0$ and, hence, $\|h^*|D_\mathcal{S} - h_\Sigma\|_\infty = 1$. □

The convergence from above established by Proposition 3.33 already suggests that we will generally have $h_N(x) > h_\Sigma(x)$ for some $x \in D_\mathcal{S}$ in the course of a CBL process. Thus, we might work with inadmissible hypotheses (see also Fig. 3.4, where $h_* \leq h_\Sigma$ does not hold). This, of course, seems to conflict with the objective of providing an outer approximation of $\varphi$. Indeed, it can easily be shown that $h_\Sigma$ is the largest function $h$ (defined on $D_\mathcal{S}$) such that $\varphi(s) \in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for all $s \in \mathcal{S}$ is guaranteed regardless of the memory $\mathcal{M}$. In other words, for each function $h$ with $h(x) > h_\Sigma(x)$ for at least one $x \in D_\mathcal{S}$, a memory $\mathcal{M}$ can be found such that $\varphi(s) \not\in \widehat{\varphi}_{h,\mathcal{M}}(s)$ for at least one $s \in \mathcal{S}$. Observe, however, that the approximation $\widehat{\varphi}_{h_N,\mathcal{M}_N}$ is derived from the *specific* memory $\mathcal{M}_N$. Thus, the fact that $h_N(x) > h_\Sigma(x)$ for some $x \in D_\mathcal{S}$ does by no means rule out the possibility of $\widehat{\varphi}_{h_N,\mathcal{M}_N}$ being an outer approximation of $\varphi$. In connection with CBLP one might therefore be interested in the probabilities

$$q_{N+1} = \mathbb{P}\left(\varphi(S_{N+1}) \not\in \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1})\right) \tag{3.28}$$

of incorrect predictions.

Consider a memory $\mathcal{M}$, a hypothesis $h$, and an input $s_0 \in \mathcal{S}$. We call $s_0$ *extremal*[23] (with respect to $\mathcal{M}$ and $h$) if $h \neq \texttt{update}(h, s_0, \mathcal{M})$, i.e., if there is some $1 \leq k \leq m$ and a case $\langle s, r\rangle \in \mathcal{M}$ such that $\sigma_\mathcal{S}(s, s_0) \in A_k$ and

$$\forall \langle s', r'\rangle \in \mathcal{M} \, : \, (\sigma_\mathcal{S}(s, s') \in A_k) \Rightarrow (\sigma_\mathcal{R}(r, r_0) < \sigma_\mathcal{R}(r, r')) \, .$$

**Lemma 3.35.** For a memory $\mathcal{M}$, a hypothesis $h \leq \text{CBLA}(\mathcal{M})$, and an input $s_0 \in \mathcal{S}$ suppose that $\varphi(s_0) \not\in \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. Then, $s_0$ is extremal. □

**Proof.** Suppose $r_0 \not\in \widehat{\varphi}_{h,\mathcal{M}}(s_0)$. Then, we find a case $\langle s, r\rangle \in \mathcal{M}$ such that $r_0 \not\in \mathcal{N}_{h(\sigma_\mathcal{S}(s,s_0))}(r)$. This means that $\sigma_\mathcal{R}(r, r_0) < h(\sigma_\mathcal{S}(s, s_0))$ and, since $h \leq \text{CBLA}(\mathcal{M})$, $\sigma_\mathcal{R}(r, r_0) < \sigma_\mathcal{R}(r, r')$ for all cases $\langle s', r'\rangle \in \mathcal{M}$ satisfying $\sigma_\mathcal{S}(s, s') \in A_{\kappa(s,s_0)}$. Hence, $s_0$ is extremal. □

**Proposition 3.36.** The following estimation holds true for the probability (3.28):

$$\begin{aligned} q_{N+1} \quad &\leq \quad \sum_{n=0}^{N} \frac{2m}{n+1} \cdot \mathbb{P}(\text{card}(\mathcal{M}_N) = n) && (3.29) \\ &\leq \quad \frac{2m}{1 + \mathbb{E}(\text{card}(\mathcal{M}_N))} = \frac{2m}{1 + \sum_{k=1}^{N} p_k}, && (3.30) \end{aligned}$$

where $m$ is the size of the partition underlying $\mathcal{H}_{step}$ and $\mathbb{E}$ denotes the expected value operator. □

**Proof.** Suppose $M_N$ to consist of $n \leq N$ cases, i.e., $\mathcal{M}_N$ is defined by some random (sub-)sequence $(S_{\pi(1)}, \ldots, S_{\pi(n)})$ of inputs, where $1 \leq \pi(1) < \pi(2) < \ldots < \pi(n) \leq N$. Moreover, consider a new input $S_0 = S_{N+1}$ and observe that

$$\mathbb{P}(\varphi(S_0) \not\in \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_0)) \leq \mathbb{P}(\varphi(S_0) \not\in \widehat{\varphi}_{h_{\mathcal{M}_N},\mathcal{M}_N}(S_0)),$$

[23] This definition of being extremal is to some extent related to the concept of "strangeness" of an observation in the context of so-called confidence machines [162, 301].

where $h_{\mathcal{M}_N} = \text{CBLA}(\mathcal{M}_N)$. From the random sequence $(S_{\pi(1)}, \ldots, S_{\pi(n)}, S_0)$ of inputs we can choose a set $\mathcal{M}'$ of (at most) $2m$ inputs resp. associated cases such that $\text{CBLA}(\mathcal{M}_N \cup \{\langle S_0, \varphi(S_0)\rangle\}) = \text{CBLA}(\mathcal{M}')$. Obviously, $\langle S_0, \varphi(S_0)\rangle \notin \mathcal{M}'$ implies that $S_0$ is not extremal. Now, recall that inputs are independent and identically distributed according to $\mu_{\mathcal{S}}$. Thus, the value $2m/(n+1)$ defines an (upper) bound to the probability that $\langle S_0, \varphi(S_0)\rangle \in \mathcal{M}'$ due to reasons of symmetry. We hence obtain

$$\begin{aligned}&\mathbb{P}(\varphi(S_0) \notin \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_0) \,|\, \text{card}(\mathcal{M}_N) = n) \leq \\ &\qquad \mathbb{P}(\varphi(S_0) \notin \widehat{\varphi}_{h_{\mathcal{M}_N},\mathcal{M}_N}(S_0) \,|\, \text{card}(\mathcal{M}_N) = n) \leq 2m/(n+1)\end{aligned}$$

from Lemma 3.35. Then, (3.29) and (3.30) follow from the theorem of total probability and Jensen's inequality, respectively. □

**Corollary 3.37.** Suppose $p \geq \delta > 0$. Then, $q_{N+1} \leq 2m/(\delta N + 1)$. Particularly, $q_{N+1} \leq 2m/(N+1)$ if $p \equiv 1$. □

According to the above results, the probability of an incorrect prediction becomes small for large memories, even though the hypotheses $h_N$ might be inadmissible. Under the assumptions of Corollary 3.37, this probability tends toward 0 with a convergence rate of order $O(1/N)$.

**Corollary 3.38.** Suppose $p \geq \delta > 0$. Then, the expected proportion of incorrect predictions in connection with CBLP converges toward 0. □

**Proof.** Define the random variable $V_n$ $(n \geq 1)$ by means of $V_n = 1$ if the $n$-th prediction is incorrect, i.e., if $\varphi(S_n) \notin \widehat{\varphi}_{h_{n-1},\mathcal{M}_{n-1}}(S_n)$, and $V_n = 0$ otherwise. Then, $\mathbb{E}(V_n) = q_n$, where $\mathbb{E}(V_n)$ denotes the expected value of $V_n$, and

$$\begin{aligned}\mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} V_n\right) &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}(V_n) \\ &\overset{\text{(Cor. 3.37)}}{\leq} \frac{1}{N}\sum_{n=1}^{N} 2m/(\delta n) \\ &\leq \frac{2m(1+\ln(N))}{\delta N} \to 0\end{aligned}$$

as $N \to \infty$. □

Example 3.39. Fig. 3.5 shows the optimal hypothesis $h^*$ for the setup $\Sigma_1$ defined in Example 2.5 (cf. Section 2.4.1) and the hypothesis $h_{\mathcal{M}}$ for a typical memory $\mathcal{M}$ of size 250, generated by a sequence of inputs chosen at random. The underlying partition has been defined by the values $\alpha_k = k/10$ $(k = 0, \ldots, 10)$. The same figure shows a characterization of the evolution of the approximation quality in the form of the values $\|h^* - h_{\mathcal{M}_n}\|_2$ and $\|h^* - h_{\mathcal{M}_n}\|_\infty$ $(n = 1, \ldots, 300)$, where $\|\cdot\|_p$ denotes the corresponding $\mathcal{L}^p$-norm. □

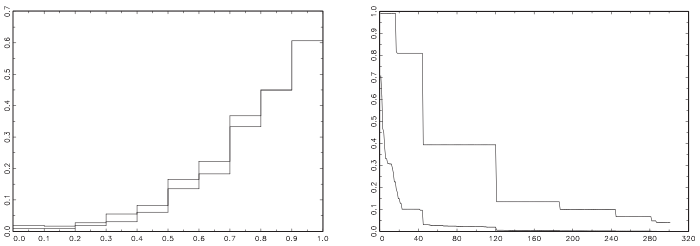


**Fig. 3.5.** Left: Optimal hypothesis $h^*$ for the setup $\Sigma_1$ in Example 2.5 and the hypothesis $h_{\mathcal{M}}$ for a memory $\mathcal{M}$ of size 250. Right: Evolution of approximation quality $\|h^* - h_{\mathcal{M}_n}\|_2$ and $\|h^* - h_{\mathcal{M}_n}\|_\infty$ (cf. Example 3.39).

The upper bound established in Proposition 3.36 might suggest to reduce the probability of an incorrect prediction by reducing the size $m$ of the partition underlying $\mathcal{H}_{step}$. Observe, however, that this will also lead to a less precise approximation of $h_\Sigma$ and, hence, to less precise predictions of outcomes. "Merging" two neighbored intervals $A_k$ and $A_{k+1}$, for instance, means to define a new hypothesis $h$ with $h|(A_k \cup A_{k+1}) \equiv \min\{\beta_k, \beta_{k+1}\}$. In fact, the probability of an incorrect prediction can be made arbitrarily small by increasing the size of the memory. The precision of the predictions, however, is limited by the precision to which $h_\Sigma$ can be approximated by $h^*$ and, hence, by the granularity of the partition underlying the definition of the hypothesis space $\mathcal{H}_{step}$. Of course, nothing prevents us from extending our approach to CBL such that it allows for the adaptation of the partition. A refinement of the latter will make sense, e.g., if the size of the memory becomes large.

Let us now consider the *fixed memory-model*, i.e., the case where CBI is based on a fixed memory $\mathcal{M} = (c_1, \ldots, c_n)$ of size $n \geq 1$. The objective of CBL is then to find an approximation of the $\mathcal{M}$-similarity profile $h_\Sigma^{\mathcal{M}}$. Thus, the consistency principle (3.22) should hold true for $\mathcal{C} = \mathcal{M} \times \mathcal{D}$. Again, the class $\mathcal{H}_*$ consists of the uncertainty minimizing hypotheses in $\mathcal{H}_{\mathcal{C}}$. Likewise, $\mathcal{H}^*$ is made of those uncertainty minimizing hypotheses that satisfy (3.22) for $\mathcal{M} \times \mathcal{D}^*$. Observation 3.30 does obviously remain correct. The hypothesis $h_* = \mathrm{CBLA}_{\mathcal{M}}(\mathcal{D})$ is now defined by the values

$$\beta_k = \min \left\{ \sigma_{\mathcal{R}}(r, r') \,|\, \langle s, r \rangle \in \mathcal{M}, \langle s', r' \rangle \in \mathcal{D}, \sigma_{\mathcal{S}}(s, s') \in A_k \right\}.$$

Thus, given a new observation, the update of the current hypothesis is realized by passing the iteration (3.26) for the $n$ cases in $\mathcal{M}$. The fixed-memory version of CBLP, denoted $\mathrm{CBLP}_{\mathcal{M}}$, is outlined in Algorithm 2.

For the hypotheses $h_N$ induced by $\mathrm{CBLP}_{\mathcal{M}}$ we do not only obtain an upper approximation but even $h_N = \mathrm{CBLA}_{\mathcal{M}}(\mathcal{D}_N)$.

**Algorithm 2** CBLP$_{\mathcal{M}}$

```
Input: a sequence of query inputs
Output: a sequence of estimation for outputs
$h_0 = \mathrm{CBLA}(\mathcal{M})$
$N = 0$
repeat
    compute $\widehat{r}_{N+1} = \widehat{\varphi}_{h_N,\mathcal{M}}(s_{N+1})$
    solve-problem($s_{N+1}, \widehat{r}_{N+1}$)
    $h_{N+1} =$ update($h_N, c_{N+1}, \mathcal{M}$)
    $N = N + 1$
until no more queries exist
```

**Proposition 3.40.** For the sequence $(h_N)_{N \geq 1}$ induced by CBLP$_{\mathcal{M}}$ it holds true that $h_N \searrow h^*$ stochastically as $N \to \infty$, where $h^*$ is defined by the values $\beta_k^* = \inf\{h_\Sigma^{\mathcal{M}}(x) \,|\, x \in D_{\mathcal{S}} \cap A_k\}$ $(1 \leq k \leq m)$. □

**Proposition 3.41.** In connection with the fixed memory-model we obtain the estimation $q_{N+1} \leq 2m/(N+1)$ for the probability (3.28), where $m$ is the size of the partition underlying $\mathcal{H}_{step}$. □

**Proof.** Consider the random sequence $(S_1, \ldots, S_N, S_0)$ of $N+1$ inputs. From this sequence we can choose a set $\mathcal{D}$ of (at most) $2m$ inputs resp. associated cases such that $\mathrm{CBLA}_{\mathcal{M}}(\mathcal{D}_N \cup \{\langle S_0, \varphi(S_0)\rangle\}) = \mathrm{CBLA}_{\mathcal{M}}(\mathcal{D})$. Now, recall that $\langle S_0, \varphi(S_0)\rangle \notin \mathcal{D}$ implies that $S_0$ is not extremal with respect to $h_N$ and $\mathcal{M}$ and that inputs are independent and identically distributed according to $\mu_{\mathcal{S}}$. Thus, the value $2m/(N+1)$ defines an (upper) bound to the probability that $\langle S_0, \varphi(S_0)\rangle \in \mathcal{D}$ due to reasons of symmetry. The rest follows from Lemma 3.35. □

**Corollary 3.42.** The expected proportion of incorrect predictions in connection with CBLP$_{\mathcal{M}}$ converges toward 0. □

It should be noticed that CBLP$_{\mathcal{M}}$ is closely related to CBLP in the case where some $N_0 \in \mathfrak{N}$ exists such that $p_N = 0$ for all $N \geq N_0$. Suppose for instance, that $p_N = 1$ for $1 \leq N < N_0$ and $p_N = 0$ for $N \geq N_0$. Then, Proposition 3.41 remains correct with CBLP$_{\mathcal{M}}$ replaced by CBLP. Proposition 3.40 remains correct if, moreover, $h_\Sigma^{\mathcal{M}}$ is replaced by $h_\Sigma^{\mathcal{M}_{N_0-1}}$. The result of Proposition 3.41 can also be used for deriving the following generalizations of Proposition 3.36 and Corollary 3.38.

**Proposition 3.43.** Let $N_0 \in \mathfrak{N}$ and suppose $p_N = 1$ for $N \geq N_0$. We then obtain the estimation

$$q_{N+1} \leq 2m \left(1 + \max\{0, N - N_0\} + \sum_{k=1}^{N_0-1} p_k\right)^{-1},$$

where $m$ is the size of the partition underlying $\mathcal{H}_{step}$. □

**Corollary 3.44.** Let $N_0 \in \mathfrak{N}$ and suppose $p_N = 1$ for $N \geq N_0$. Then, the expected proportion of incorrect predictions in connection with CBLP converges toward 0. □

Summing up, the results of this section throw light on some interesting properties of our approach to case-based learning. In fact, the combination of case-based inference and case-based learning, i.e., the application of the prediction scheme of Section 3.2.1 with a hypothesis derived by means of CBLA, allows for deriving a set-valued prediction $\widehat{\varphi}(s_0) = \widehat{\varphi}_{h,\mathcal{M}}(s_0)$ which covers the true outcome with a high probability. In a statistical sense, $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ can thus be seen as a kind of confidence region or *credible output set*, a justification for designating the above inference scheme as *credible case-based inference*.

Remark 3.45. In many applications one is interested in both, a credible output set and a "point-estimation" of the output $r_0$, i.e., a distinguished element $\hat{r}_0 \in \mathcal{R}$ that can be considered as representative. The latter can be derived from the credible output set $\widehat{\varphi}_{h,\mathcal{M}}(s_0)$ as a *generalized median*:

$$\hat{r}_0 \overset{\mathrm{df}}{=} \arg \max_{r \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \sum_{r' \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)} \sigma_{\mathcal{R}}(r, r') \tag{3.31}$$

As can be seen, the generalized median is a kind of center-point, namely the element of the credible output set which is maximally similar to all other elements. □

Note that the concrete probability of a correct prediction depends on the number of observed cases and can thus be estimated in advance. Moreover, it can be made arbitrarily large by extending the size of the memory. CBLP, the combination of CBI and CBL, can thus be seen as an interesting method of statistical inference. Principally, it defines a generalized instance-based learning algorithm which takes uncertainty in connection with the prediction of outcomes into account. This aspect will be discussed in more detail in Section 3.5 below.

Let us finally mention that results similar to the ones derived in this section can also be obtained in connection with other types of similarity profiles. Recall, for instance, the concept of a *local* similarity profile: Let $\mathcal{M}$ be a memory of cases, namely a subset $\mathcal{M} \subseteq \mathcal{D}$ of the cases $\langle s_n, r_n \rangle$ $(1 \leq n \leq N)$ which have been encountered so far. For $\langle s, r \rangle \in \mathcal{M}$ we define the *local hypothesis* $h^s$ by the values

$$\beta_k \overset{\mathrm{df}}{=} \min_{1 \leq n \leq N : \sigma_{\mathcal{S}}(s, s_n) \in A_k} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s_n)). \tag{3.32}$$

The *local* $\mathcal{M}$-*hypothesis* is given by $h^{\mathcal{M}} \overset{\mathrm{df}}{=} \{h^s \,|\, s \in \mathcal{M}^{\downarrow}\}$. We can then prove a result similar to Proposition 3.36:

**Proposition 3.46.** Suppose that $N$ (independent and identically distributed) cases have been encountered so far. For a subset $\mathcal{M}$ containing $|\mathcal{M}|$ cases let a local $\mathcal{M}$-hypothesis be defined according to (3.32). Moreover, let $s_0 \in \mathcal{S}$ be a new problem (chosen at random from $\mathcal{S}$). The probability that the true outcome $r_0 = \varphi(s_0)$ is not covered by

$$\widehat{\varphi}_{h^{\mathcal{M}},\mathcal{M}}(s_0) = \bigcap_{\langle s,r\rangle \in \mathcal{M}} \mathcal{N}_{h^s(\sigma_{\mathcal{R}}(s,s_0))}(r) \tag{3.33}$$

is bounded from above by $|\mathcal{M}|m/(N+1)$. □

A prediction (3.33) based on a local $\mathcal{M}$-hypothesis is generally more precise than a prediction (3.2). At the same time, however, the associated confidence level is smaller. Still, Proposition 3.46 shows that this level can be made arbitrarily large by increasing the number of observed cases.

Note that it might not be possible to compute the hypothesis (3.32) exactly if only some of the encountered cases $\langle s_n, r_n \rangle \in \mathcal{D}$ are added to $\mathcal{M}$. However, Proposition 3.46 remains valid (up to some minor modifications) if the minimum in (3.32) is not taken over all (pairs) of cases.

### 3.4.4 Experimental results

The basic learning scheme presented in Section 3.4.2 offers a convenient framework which enables the realization of methods for predicting unknown outcomes based on a sequence of observed cases. The results of Section 3.4.3 show that corresponding predictions take the form of confidence regions which cover the unknown output with a certain probability. In this section, we shall present some small examples in order to convey how this approach works in practice. These examples are not meant as an empirical evaluation of our CBI method, they are only intended to provide an illustration of the theoretical results derived above.

We have organized two experimental studies as follows: First of all, a target function $\varphi$ with domain $\mathcal{S}$ and range $\mathcal{D}$ is specified. A single run of a simulation corresponds to the CBLP scheme presented in Section 3.4.3, where $p \equiv 1$, a new input is chosen according to the uniform distribution, and the length of the generated random sequence of inputs is 1000. The size of the partition underlying the learned similarity hypothesis is $m = 20$. Given a new input $S_{N+1}$, a prediction $\widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1})$ is derived from the hypothesis $h_N$ and the memory $\mathcal{M}_N$ according to (3.15) or (3.16). Two characteristic quantities are recorded for this estimation. Firstly, the *correctness* is captured by means of $V_N \in \{0,1\}$, where $V_N = 1$ iff $(\varphi(S_{N+1}) \in \widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1}))$. Secondly, the *precision* is specified by $P_N \overset{\mathrm{df}}{=} \mathrm{diam}(\widehat{\varphi}_{h_N,\mathcal{M}_N}(S_{N+1}))$. The behavior of the prediction method can then be characterized by means of the expected values $\mathbb{E}(V_N)$ and $\mathbb{E}(P_N)$ associated with the sequences $(V_1, \ldots, V_{1000})$ and $(P_1, \ldots, P_{1000})$, respectively. Approximations of these expected values have been obtained by deriving mean values $\overline{V}_N$ and $\overline{P}_N$ from a large number of simulation runs. The respective sequences $(\overline{V}_1, \ldots, \overline{V}_{1000})$ and $(\overline{P}_1, \ldots, \overline{P}_{1000})$ constitute the results which are finally presented in Appendix D. Note that $1 - \overline{V}_N$ is an estimation of the probability $q_N$ specified in (3.28).

For the first example, we have chosen the relatively simple function

$$\varphi : s \mapsto \sin(s+1) \cdot \cos^2(s),$$

where $\mathcal{S} = [0, \pi/2] \cap \mathfrak{Q}$ (and $\mathcal{R} = \varphi(\mathcal{S}) \subseteq [0, 1.2]$). The results are shown in Fig. D.1–Fig. D.3. As it was to be expected from the theoretical results of Section 3.4.3, the probability of an incorrect prediction soon becomes very small. Of course, the more cases are used for constraining the outcome, the more precise the predictions become. At the same time, however, this also increases the probability of an incorrect prediction. The approximation (3.16), using a constant number of $k = 10$ cases, shows that the expected precision of a prediction is not necessarily a monotone function of the size of the memory (cf. Fig. D.2). This effect is not restricted to (3.16) but can also occur in connection with (3.15), i.e., if all cases are used. It is caused by two opposite effects related to the extension of a memory. On the one hand,

$$\mathcal{M}' \subseteq \mathcal{M} \;\Rightarrow\; \widehat{\varphi}_{h,\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h,\mathcal{M}'}(s)$$

for all hypotheses $h$, memories $\mathcal{M}, \mathcal{M}'$, and $s \in \mathcal{S}$. That is, the larger a memory is, the more precise the approximation becomes. On the other hand,

$$h \leq h' \;\Rightarrow\; \widehat{\varphi}_{h',\mathcal{M}}(s) \subseteq \widehat{\varphi}_{h,\mathcal{M}}(s)$$

for all hypotheses $h, h'$, i.e., the less strong a hypothesis is, the less precise the approximation becomes. The aforementioned effect is then explained by the fact that a case-based approximation is derived from a memory $\mathcal{M}$ and the associated hypothesis $h_{\mathcal{M}}$ and that $\mathcal{M}' \subseteq \mathcal{M}$ implies $h_{\mathcal{M}} \leq h_{\mathcal{M}'}$.

The simulation results might give the impression that the expected precision of predictions converges toward some value which is larger than 0. Even though this might happen in certain cases, it is actually not true for our example. In fact, this example reflects a typical situation where the expected precision indeed converges toward 0, but where the improvement due to additional observations decreases with the size of the memory. In other words, the convergence rate might be rather low. This can also be illustrated by means of the simple example $\varphi : s \mapsto s^2$, $s \in \mathcal{S} = [0, 1]$.[24] For the CBI setup using $\sigma_{\mathcal{S}} : (s, s') \mapsto 1 - |s - s'|$ and $\sigma_{\mathcal{R}} : (r, r') \mapsto 1 - |r - r'|$ we obtain $h_{\Sigma}(x) = x^2$. Moreover, it can be shown that (3.15) leads to $\widehat{\varphi}_{h_{\Sigma},\mathcal{M}}(0) = [0, 2\min\{s_1, \ldots, s_n\}]$, where $s_1, \ldots, s_n$ denote the inputs which have already been observed, i.e., which define the memory $\mathcal{M}$. That is, the expected precision of the prediction of $\varphi(0)$, i.e., the length of the above interval, is given by the random variable $X \overset{\mathrm{df}}{=} 2\min\{S_1, \ldots, S_n\}$,

[24] For the sake of simplicity, we put up with the fact that $\mathcal{S}$ violates our assumption of countability.

where $S_1, \ldots, S_n$ are independent random variables distributed according to $\mu_{\mathcal{S}}$. If the latter is taken as the uniform measure over $[0,1]$, it is not difficult to show that $\mathbb{E}(X) = 2/(n+1)$. Thus, the expected precision converges toward 0 with a convergence rate of $O(1/n)$.

The second experimental study uses the CBI setup $\Sigma_1$ which has been introduced in Example 2.5, i.e., a value $\varphi(s)$ is defined as the cost of the optimal solution associated with the combinatorial optimization problem encoded by $s$. The results of this study, shown in Fig. D.4–Fig. D.7, are qualitatively similar to those of the first experiment. As can be seen in Fig. D.4, the non-monotone behavior of the expected precision of predictions now also occurs in connection with the case-based approximation (3.15). It should be remarked that the results are quite satisfactory in the sense that a rather small fraction of the $\mathrm{card}(\mathcal{S}) = 7^5$ cases suffices for deriving relatively precise predictions of cost values (which are between 0 and 48). A memory of size 1000, for instance, corresponds to a fraction of approximately 6/100, i.e., a prediction based on the 10 most similar cases uses only slightly more than 0.06% of the cases.

Let us finally consider a "real-world" application. In connection with the HOUSING DATABASE,[25] we have used CBI for predicting prices of houses which are characterized by 13 attributes. Similarity was defined as an affine-linear function of the distance between (real-valued) attribute values. For randomly chosen memories of size 30 we have used 450 cases as training examples in order to learn the respective local $\mathcal{M}$-profiles. Based on (local) hypotheses thus obtained, CBI allowed for predicting prices of the remaining 56 cases with a precision of approximately 10,000 dollars and a confidence level around 0.85. Taking the generalized median (3.31) as a point-estimation, which here simply corresponds to the center of the interval, one thus obtains predictions of the form $x \pm 5,000$ dollars. As can be seen, these estimations are quite reliable but not extremely precise (the average price of a house is approximately 22,500 dollars). In fact, this example clearly points out the limits of an inference scheme built upon the CBI hypothesis. Our approach takes these limits into account and makes them explicit: A case-based prediction of prices cannot be confident and extremely precise at the same time, simply because the housing data meets the CBI hypothesis but moderately. Needless to say, problems of such type are of a general nature and by no means specific to case-based inference. Linear regression, for example, assumes a linear relationship between the dependent and independent variables. It yields poor predictions and imprecise confidence intervals if this assumption is not satisfied (which is often the case in practice).

[25] Available at `http://www.ics.uci.edu/~mlearn`.

## 3.5 Application to statistical inference

It has already been mentioned that our approach to case-based learning (Section 3.4) gives rise to an extension of the inference scheme of Section 3.2 which provides us with an interesting statistical inference mechanism. In fact, it is just the attached level of confidence which makes a (set-valued) prediction (3.2) attractive from a statistical perspective. In order to emphasize this point, we have already used the term *credible* CBI, referring to the combination of the inference scheme (3.2) and the case-based learning algorithm of Section 3.4: Given a randomly chosen memory $\mathcal{M}$ of cases and a new input $s_0$, CBI derives a hypothesis $h = \mathrm{CBLA}(\mathcal{M})$ and delivers a prediction

$$(\widehat{\varphi}_{h,\mathcal{M}}(s_0), \alpha)$$

such that

$$\mathbb{P}\left(\varphi(s_0) \in \widehat{\varphi}_{h,\mathcal{M}}(s_0)\right) \geq 1 - \alpha.$$

This section is meant to outline briefly two applications which show that credible CBI can complement existing statistical methods in a reasonable way.

### 3.5.1 Case-based parameter estimation

In order to show how credible CBI might support classical approaches to statistical inference let us consider the idea of *case-based parameter estimation.* Thus, the task is to estimate an unknown parameter $\vartheta \in \Theta$, where $\Theta$ denotes an underlying class of parameters. Quite often, the estimation of $\vartheta$ according to, say, the Maximum Likelihood (ML) principle, is a computationally complex problem involving numerical optimization methods. The computation of an ML estimation (MLE) is hence impossible if such estimations have to be made available frequently, perhaps even under strict time constraints. As an example one might think of a control problem where data is obtained from monitoring a technical system and where the MLE serves as a control parameter [219]. Likewise, online data analysis and estimation problems arise in mining so-called *data streams* [92, 161].

If the (repeated) derivation of an MLE is computationally too complex, credible CBI might be used for estimating it. More specifically, we can derive a confidence region for the MLE based on a set of data–MLE tuples and a new set of data. Using our terminology, the data plays the role of an input and the MLE corresponds to the output. The data–MLE tuples which constitute the memory may originate from other estimations or may have been derived during a less time-critical preprocessing phase.

The CBI hypothesis now means that similar data leads to similar ML estimations, an assumption which appears reasonable for many applications. Still, the choice of an adequate measure for determining the similarity between two sets of data will generally not be obvious. Since the adequacy of a measure depends on the respective application, we will not go into detail here. Let us only mention that it will often be possible to simplify the problem by passing from the data itself to *sufficient statistics* thereof, i.e., to consider sufficient statistics as inputs which determine the output in the form of an MLE.

In general, one will be interested in a confidence region not for the MLE $\vartheta_{ML}$ but for the *true* parameter $\vartheta$ of an underlying stochastic model. Suppose that a confidence region for $\vartheta$ takes the form $\vartheta_{ML} \oplus C_{ML}$, where $C_{ML} \subseteq \mathfrak{R}^n$ can be constructed from the data and does not depend on $\vartheta$. A simple example is the estimation of the mean $\mu$ of a normal distribution with standard deviation $\sigma$. In this case, the $(1-\alpha)$-confidence region $C_{ML}$ corresponds to an interval $[-t_\alpha \cdot \sigma/\sqrt{n}, t_\alpha \cdot \sigma/\sqrt{n}]$,[26] i.e., $C_{ML}$ depends only on the number of observations. Now, let $(\widehat{\varphi}_{h,\mathcal{M}}, \beta)$ be a CBI prediction of $\vartheta_{ML}$. Since

$$(\vartheta_{ML} \in \widehat{\varphi}_{h,\mathcal{M}}) \wedge (\vartheta \in \vartheta_{ML} \oplus C_{ML}) \Rightarrow (\vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \oplus C_{ML}),$$

we obtain

$$\mathbb{P}(\vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \oplus C_{ML}) \geq (1-\alpha)(1-\beta).$$

That is, the set $\widehat{\varphi}_{h,\mathcal{M}} \oplus C_{ML}$ defines a $(1-\alpha)(1-\beta)$-confidence region for $\vartheta$. This way, a confidence region for the true parameter $\vartheta$ can be derived by means of purely *case-based* reasoning, i.e., without any reference to a likelihood function and corresponding maximization problems.

### 3.5.2 Case-based prior elicitation

The determination of prior probability distributions is a main burden of Bayesian analysis, and it has become a focus of criticism of the Bayesian approach. As a second application let us therefore consider the possibility of exploiting (credible) CBI in order to support the elicitation of such priors, i.e., the determination of prior distributions from previous cases. The idea is thus to treat a CBI prediction $(\widehat{\varphi}_{h,\mathcal{M}}, \alpha)$ of an MLE $\vartheta_{ML}$ as prior information about the unknown parameter $\vartheta$.

In general, there will exist several possibilities of utilizing a CBI prediction. A relatively straightforward choice of a prior based on a prediction $(\widehat{\varphi}_{h,\mathcal{M}}, \alpha)$ is defined by the associated probability density function

$$f : \vartheta \mapsto \left\{ \begin{array}{cl} (1-\alpha)(\int_{\widehat{\varphi}_{h,\mathcal{M}}} dt)^{-1} & \text{if} \quad \vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \\ \alpha(\int_{\Theta \setminus \widehat{\varphi}_{h,\mathcal{M}}} dt)^{-1} & \text{if} \quad \vartheta \not\in \widehat{\varphi}_{h,\mathcal{M}} \end{array} \right. ,$$

where we assume $(\int_\Theta dt) < \infty$.[27] For very small $\alpha$ one might even completely concentrate on the predicted region and define a corresponding uniform prior only over $\widehat{\varphi}_{h,\mathcal{M}}$:

[26] The value $t_\alpha$ is defined through the equality $\int_{-t_\alpha}^{t_\alpha} \phi(t)\,dt = 1-\alpha$, where $\phi$ denotes the probability density function of the standard normal distribution.

[27] Otherwise it might still be possible to work with improper priors.

$$f : \vartheta \mapsto \begin{cases} (\int_{\widehat{\varphi}_{h,\mathcal{M}}} dt)^{-1} & \text{if } \vartheta \in \widehat{\varphi}_{h,\mathcal{M}} \\ 0 & \text{if } \vartheta \notin \widehat{\varphi}_{h,\mathcal{M}} \end{cases} .$$

The prior distribution is often assumed to belong to a certain parameterized class $\mathcal{C} = \{f_\gamma \,|\, \gamma \in \Gamma\}$ of distributions, where $\mathcal{C}$ is chosen in such a way that the prior is *conjugate* to the likelihood function. This guarantees that the posterior distribution belongs to the same class. A CBI prediction can then be utilized for constraining (or even determining) the parameters of a prior distribution, the so-called *hyper-parameters*. More precisely, a prediction $(\widehat{\varphi}_{h,\mathcal{M}}, \alpha)$ serves as a constraint in the sense that the parameter $\gamma$ has to satisfy $\int_{\widehat{\varphi}_{h,\mathcal{M}}} f_\gamma(t)\, dt = \alpha$. For example, if the prior is normal with mean $\mu$ and standard deviation $\sigma$, the CBI prediction $([\beta^-, \beta^+], \alpha)$ entails $\int_{\beta^-}^{\beta^+} \phi_{\mu,\sigma}(t)\, dt = \alpha$, which in turn suggests

$$\mu = \frac{\beta^- + \beta^+}{2}, \qquad \sigma = \frac{\beta^+ - \beta^-}{2t_\alpha}.$$

## 3.6 Summary and remarks

### Summary

- We have adopted a *constraint-based* view of the CBI hypothesis, according to which the similarity of inputs imposes a constraint on the similarity of associated outcomes in the form of a lower bound. This interpretation allows for exploiting the reasoning principle underlying CBI within a formal inference process.
- The concept of a *similarity profile* has been introduced. It establishes a connection between the system level and the similarity level and represents the similarity structure of a CBI setup. Several generalizations of this concept have been proposed in order to take special characteristics of CBI into consideration and to improve case-based inference.
- A *similarity hypothesis* is thought of as an approximation of a similarity profile. It thus defines a formal model of the CBI hypothesis for the system under consideration.
- CBI has been realized as a process of constraint propagation which allows for predicting an unknown output $r_0 \in \mathcal{R}$ by means of a set $\widehat{\varphi}_{h,\mathcal{M}}(s_0) \subseteq \mathcal{R}$ of possible outcomes. This set is derived from an underlying hypothesis $h$ and a memory $\mathcal{M}$ of cases. It is guaranteed to cover $r_0$ if $h$ is admissible. An efficient implementation of this inference scheme can be realized by means of parallel computation techniques.
- We have studied some properties of *case-based approximations*, i.e., set-valued mappings $\widehat{\varphi}_{h,\mathcal{M}} : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$ derived from a hypothesis $h$ and a memory of cases $\mathcal{M}$.

- The idea of *case-based learning* can be realized in different ways within our framework. Here, we have concentrated on the learning of a suitable similarity hypothesis from a sequence of observations.
- Utilizing the hypothesis space $\mathcal{H}_{step}$, which consists of a class of step functions on $[0, 1]$, allows for realizing CBL by means of an efficient candidate-elimination algorithm, CBLA. Particularly, the time complexity of updating a hypothesis $h_{\mathcal{M}}$ is linear in the size of the memory $\mathcal{M}$.
- A sequence of hypotheses derived by CBLA from a random sequence of cases converges stochastically toward the optimal admissible hypothesis $h^* \in \mathcal{H}$. Even though these hypotheses may be inadmissible, they allow for deriving predictions which define outer bounds with high probability. We thus obtain a method of *credible case-based inference* that produces predictions in the form of *credible output sets* which cover the true output with high probability. In fact, our CBI method can be seen as a non-parametric approach to estimating confidence regions.
- In Section 3.5, it has been argued that credible CBI is also interesting in the context of classical statistical inference. More specifically, we have outlined the ideas of case-based parameter estimation and case-based prior elicitation in Bayesian analysis.

### Remarks

- Within our framework, the concept of *similarity* should be seen as an essential but at the same time *auxiliary* concept. Indeed, the inference procedure outlined in this chapter principally works with *any* pair of similarity functions $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$, each of which defines a certain similarity structure. Of course, the more suitably these functions are chosen, the more precise the inference results will be. However, since our inference scheme takes into account the degree to which the CBI hypothesis applies these results remain valid even if similarity is not quantified in a meaningful way. The interpretation as an auxiliary concept contrasts with other formalizations of CBI [99, 141, 296], in which inference becomes more or less meaningless without a reasonable measure of similarity.
- It has already been remarked that the CBI scheme in Section 3.2.1 is based on the transformation of original data, i.e., instances in the space $\mathcal{S} \times \mathcal{R}$, into points of the similarity space $D_{\mathcal{S}} \times D_{\mathcal{R}}$. In this connection, it is interesting to note that the transformation of data from a high-dimensional into a low-dimensional space is also used by several other methods, e.g., in statistical data analysis or self-organizing neural networks. Of course, the underlying objective which is common to these methods is to capture essential properties of a system structure by means of a simplified representation.
- In [221], an instance-based prediction method has been advocated as an alternative to linear regression techniques. By deriving set-valued instead of point

estimations, credible CBI somehow combines advantages from both methods: It requires less structural assumptions than (parametric) statistical methods as does the instance-based approach. Still, it allows for quantifying the uncertainty related to predictions by means of confidence regions. We shall return to this point in the following chapter.

- We have argued that our approach to CBI combines model-based and instance-based learning (cf. Section 3.1). Let us mention, therefore, another idea of establishing a relationship between model-based and instance-based reasoning. According to the point of view adopted in [236], an instance-based prediction is obtained within a Bayesian framework by marginalizing over all possible model families and all (parameterized) individual models within those families. The basic idea can be expressed by writing (in a somewhat sloppy notation)

$$\mathbb{P}(x \,|\, X) = \int_{\mathcal{M}} \mathbb{P}(x \,|\, M)\, \mathbb{P}(M \,|\, X)\, dM, \tag{3.34}$$

  where $X$ and $x$ denote, respectively, the observed data and a new vector $x$, and $\mathcal{M}$ is a class of models. Equation (3.34) suggests that the prediction does not depend on a model, only on the data $X$. However, apart from some technical difficulties, this approach is not very convincing. In fact, (3.34) is nothing else than the standard approach to higher-level Bayesian analysis (Bayesian averaging): A prediction is derived by taking the average of the predictions made by each possible model, weighted by the plausibilities of these models. Thus, (3.34) corresponds to a weighted average of models of a certain class (sometimes called the ensemble average).[28] It is by no means "model-free" since the bias of the model class is actually not "integrated out" by (3.34). Besides, it deserves mentioning that our approach to combining model-based and instance-based inference is very different. This becomes especially obvious by realizing that we do not consider a model of any underlying data-generating process, but rather of the CBI principle itself.

- The construction of confidence regions[29] in the context of CBLP is in line with classical (Neyman-Pearson) statistical inference. Particularly, the inference procedure does not condition on the (structure of the) actually observed data (as likelihood methods do). Rather, the claim that the $n$-th outcome is covered with probability $1-\alpha_n$ by the confidence region $C_n$ derived from the first $n-1$ cases should be interpreted in a "frequentistic" way: Let an experiment consist of drawing a random sample of $n$ cases, constructing a confidence region from the first $n-1$ cases, and noting a success if the outcome of the $n$-th case is covered by that region. By repeating this type of experiment over and over again, the relative frequency of successes will converge toward $1-\alpha_n$. In other words, the probability $\alpha_n$ is a property which has to be ascribed to the *inference procedure*, not to the result.

[28] Taking all model families (whatever this means) into account is impossible anyway. In practice, one only considers one class, e.g., a certain type of neural networks.

[29] Note that these regions are random variables.

- In Section 3.1, we have hinted at limitations of a similarity-based analysis which can occur due to the low dimensionality of the similarity space. In order to overcome such limits one might think of using a more general, multi-dimensional formalization of the concept of similarity. Such representations have indeed been advocated in literature (e.g. [283]).
- In [247], the authors consider the problem to quantify the extent to which the CBR hypothesis holds for a particular application at hand. To this end, they propose a measure of the *problem–solution regularity*. In contrast to our concept of a similarity profile, however, this is a one-dimensional measure. Besides, it is not used for the purpose of prediction but rather as a kind of trigger for the maintenance of the CBR system.

# 4. Probabilistic Modeling of Case-Based Inference

The main idea of case-based inference is to exploit the information provided by the *similarity structure* of a problem $\langle \Sigma, s_0 \rangle$ in order to improve the prediction of an unknown outcome $r_0 = \varphi(s_0)$. In Chapter 3, this structure has been characterized by means of the similarity profile $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$ and the similarity structure $\mathsf{SST}(\mathcal{M}, s_0)$. The specification of lower similarity bounds by $h_\Sigma$ allows for the derivation of (set-valued) predictions which are guaranteed to cover the unknown outcome. Still, $h_\Sigma$ gives only a relatively crude picture of the similarity structure of the setup $\Sigma$, and predictions thus derived may turn out to be rather imprecise.

In particular, due to the fact that a similarity profile provides worst case estimations in the form of lower similarity bounds, it is rather sensitive toward outliers, i.e., similarity pairs

$$(x, y) = (\, \sigma_\mathcal{S}(s_\imath, s_\jmath), \sigma_\mathcal{R}(r_\imath, r_\jmath)\,) \tag{4.1}$$

with comparatively small $y$. In fact, as $h_\Sigma(x)$ is a lower bound to the similarity of outputs that belong to $x$-similar inputs, even the existence of a single pair of $x$-similar inputs having rather dissimilar outcomes entails a small lower bound $h_\Sigma(x)$. Small bounds in turn will obviously have a negative effect on the precision of (set-valued) predictions (3.2). This problem is illustrated in Fig. 4.1 for the `auto-mpg` data set, a benchmark from the UCI repository.[1] The picture clearly reveals the aforementioned outlier effect: The similarity profile (similarity hypothesis) is "pressed down" by a relatively small number of similarity pairs (4.1).

Due to the above problem, credible case-based predictions will often be very imprecise. The concept of a *local* similarity profile, that has already been proposed in Chapter 3, may alleviate this problem, as the similarity bounds in local profiles only refer to a local region in the input space. An alternative or rather complementary idea is to weaken the concept of a similarity profile by looking for similarity bounds that are "almost valid", that is, valid with a certain probability. Before formalizing this idea in a rigorous way, let us introduce a simpler example that we shall use for illustration purposes throughout this chapter.

Example 4.1. Consider the $30 \times 30$ grid shown in Fig. 4.2. This grid is thought of as encoding a *fuzzy* concept or category. An input $s \in \mathcal{S}$ corresponds to an

[1] Here, the problem is to predict the fuel consumption (= outputs) of cars (= inputs) which are characterized by a number of attributes like horsepower or size. See section 4.4 for more details, including the specification of the underlying similarity measures.

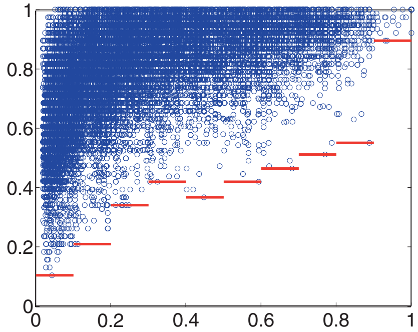


**Fig. 4.1.** Similarity hypothesis for the `auto-mpg` data (step function). Each point corresponds to a pair $(x, y)$ with $x = \sigma_{\mathcal{S}}(s_\imath, s_\jmath)$ (abscissa) and $y = \sigma_{\mathcal{R}}(r_\imath, r_\jmath)$ (ordinate).

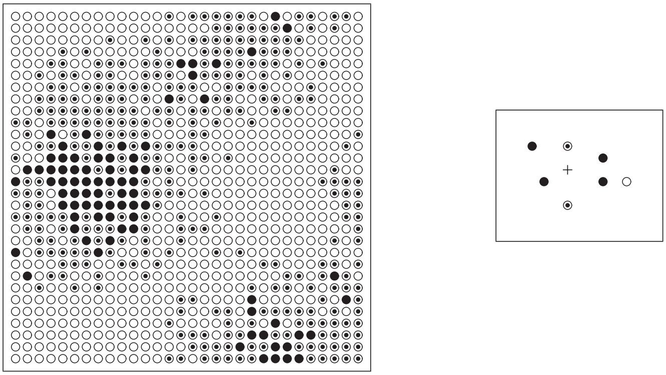


**Fig. 4.2.** Illustration of the fuzzy concept of Example 4.1 (left): A black circle corresponds to a positive example, a white circle to a negative one, and a black and white circle indicates a membership degree of 1/2. Right: Illustration of a prediction task.

*instance* and is identified by the coordinates of a grid point, i.e., $\mathcal{S} = \{(\iota, \kappa) \,|\, 1 \leq \iota, \kappa \leq 30\}$. Let $\mu_{\mathcal{S}}$ be the uniform measure over $\mathcal{S}$. The set of outputs is defined as $\mathcal{R} = \{0, 1/2, 1\}$ and encodes the degree of membership of an associated instance: $\varphi(s) = 1$ means that the instance $s$ is a *positive example* for the concept, $\varphi(s) = 0$ corresponds to the case of a *negative example*, and $\varphi(s) = 1/2$ means that $s$ belongs "more or less" to the category. Let

$$\sigma_{\mathcal{S}}\left((\iota, \kappa), (\iota', \kappa')\right) = \max\left\{0, 1 - \frac{1}{7} \cdot \max\{|\iota - \iota'|, |\kappa - \kappa'|\}\right\}.$$

Moreover, let $\sigma_{\mathcal{R}}$ be defined as $(r, r') \mapsto 1 - |r - r'|$. This setup will henceforth be referred to as $\Sigma_3$. In connection with the problem of predicting membership degrees of instances, we might take advantage of the CBI hypothesis suggesting that instances which are "close" to each other have "similar" degrees of membership. Consider, for instance, the prediction task which is also illustrated in Fig. 4.2. The observation of some of the surrounding instances will obviously have an effect on our belief concerning the membership of the new instance which is marked by a cross. However, even if intuitively justified, the CBI assumption is actually not valid, at least if taken literally: One will always find (at least) one pair of instances which are neighbored (similar) to a certain extent and for which the claim of similar membership degrees does not apply. In fact, formally we derive the lower bounds $h_{\Sigma}(x) = 0$ for all $x \in D_{\mathcal{S}} \setminus \{1\} = \{0, 1/7, 2/7, \ldots, 6/7\}$. The important point, however, is the observation that the hypothesis holds true for *most* of the examples and, hence, could still support the aforementioned prediction task. □

In this chapter, we shall approach the problem of deriving better predictions by means of probabilistic methods. An obvious approach to achieving this is to consider probability measures as a refinement of set-valued predictions. In accordance with the constraint-based approach of Section 3.2, we might thus look at the probabilities

$$\mathbb{P}\left(R_0 = r \,|\, \mathsf{OST}(\mathcal{M}, s_0)\right). \tag{4.2}$$

(4.2) specifies the probability that the outcome $\varphi(s_0)$, which is now treated as a random variable $R_0$, is realized by $r \in \mathcal{R}$, given the information provided by the outcome structure of the problem $\langle \Sigma, s_0 \rangle$. More specifically, the *indirect* character of CBI (cf. the remarks on page 67) suggests to proceed from the similarity structure of $\langle \Sigma, s_0 \rangle$ and, hence, to derive probabilities

$$\mathbb{P}\left(Y = y \,|\, \mathsf{SST}(\mathcal{M}, s_0)\right) \tag{4.3}$$

of corresponding similarity degrees $y \in (D_{\mathcal{R}})^n$ first. Evidence concerning the outcome $r_0$ is then derived in a second step from (4.3).

Remark 4.2. Treating $r_0$ as a random variable can be justified (from a "frequentist" viewpoint) even if we do not adopt a subjective (Bayesian) position. In Section 2.4, we have mentioned the idea of using a memory repeatedly for solving instances drawn at random from a certain class of (combinatorial optimization) problems. If the similarity structure of the setup $\Sigma$ is indeed informative, the probabilistic approach will make problem solving more efficient *on the average*. In fact, the idea of *repetitive* problem solving is bound up with CBI.

Note that CBI does not take information about the new input itself into account, but only about the similarity between the new and already observed inputs. Indeed, looking at $r_0$ as a random variable would otherwise become more dubious, even from a "subjectivist" point of view. Namely, for a deterministic CBI problem the output $r_0$ is actually determined as soon as $s_0$ is known, even though its derivation might involve a computationally complex process.[2] □

A further possibility of generalizing the method presented in Chapter 3 is to fall back on the idea underlying the *likelihood principle* in statistical inference and, hence, to look at the *likelihood* function [139]

$$\lambda : r \mapsto \mathbb{P}\left(\mathsf{OST}(\mathcal{M}, s_0) \,|\, R_0 = r\right). \tag{4.4}$$

In fact, the constraint-based approach of Chapter 3 can also be interpreted as a special realization of this idea: Using only the likelihood degrees 0 and 1, the likelihood of a (hypothetical) outcome $r \in \mathcal{R}$ is 0 as soon as it is not compatible with the outcome structure of a problem $\langle \Sigma, s_0 \rangle$. Besides, an approach based on (4.4) also presents the possibility of realizing a Bayesian reasoning procedure in which the likelihood function serves as evidence for updating a (probabilistic) quantification of the (prior) belief about the outcome $r_0$.

The remaining part of this chapter is organized as follows: Section 4.1 provides probabilistic generalizations of the concepts which have been introduced in Section 3.1. In Section 4.2, some general aspects concerning the relation between CBI, probabilistic reasoning and statistical inference are discussed. Section 4.3 is concerned with approaches to case-based learning within the probabilistic setting. Moreover, it proposes a generalization of the constraint-based inference scheme from the previous chapter which produces a nested sequence of credible output sets associated with different levels of confidence. Related experimental results are presented and discussed in Section 4.4. The subsequent sections are devoted to alternative types of probabilistic inference: The representation of case-based evidence in the form of belief functions and the combination of individual pieces of evidence in the framework of information fusion (Section 4.5), CBI based on more complex (probabilistic) similarity profiles (Section 4.7), and approximate probabilistic inference schemes which can be seen as a direct generalization of the constraint-based approach to CBI (Section 4.8).

## 4.1 Basic probabilistic concepts

In this section, we introduce probabilistic generalizations of the concepts which have been discussed in Section 3.1. Consider a problem $\langle \Sigma, s_0 \rangle$ with

[2] This is related to a problem discussed under the slogan "logical omniscience" by philosophically minded logicians and probabilists. See [182] for an interesting discussion and a related extension of Savage's framework of subjective probability theory.

$$\mathcal{M} = \big(\langle s_1, r_1\rangle, \ldots, \langle s_n, r_n\rangle\big) \tag{4.5}$$

being the memory of the setup $\Sigma$. According to our probabilistic modeling of the occurrence of inputs, the sequence $(s_1, \ldots, s_n, s_0)$ can be seen as the realization of a random sequence of inputs which is characterized by the probability measure

$$(\mu_{\mathcal{S}})^{n+1} \overset{\mathrm{df}}{=} \underbrace{\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}} \otimes \ldots \otimes \mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}}}_{n+1 \text{ times}} \in \mathcal{P}(\mathcal{S}^{n+1}). \tag{4.6}$$

This measure defines the (discrete) probability space $(\mathcal{S}^{n+1}, (\mu_{\mathcal{S}})^{n+1})$ underlying the CBI problem. It determines the probability of the occurrence of certain information structures, such as the outcome structure $\mathsf{OST}(\mathcal{M}, s_0)$. Observe that the memory (4.5) is distributed according to $\mu_{\mathcal{S}\times\mathcal{R}}$, where

$$\mu_{\mathcal{S}\times\mathcal{R}}(\{(s,r)\}) = \begin{cases} \mu_{\mathcal{S}}(\{s\}) & \text{if} \quad \varphi(s) = r \\ 0 & \text{if} \quad \varphi(s) \neq r \end{cases} .$$

In accordance with the CBI hypothesis, case-based inference is particularly concerned with modeling the (similarity) relation between *pairs* of cases. Thus, we shall pay special attention to (4.6) with $n = 1$. The more general case $n > 1$ and the related problem of combining (probabilistic) evidence obtained from different cases will be discussed in subsequent sections.

### 4.1.1 Probabilistic similarity profiles and hypotheses

Consider a random tuple $(S, S') \in \mathcal{S} \times \mathcal{S}$ of inputs. The random variable $Z = (X, Y)$, with $X = \sigma_{\mathcal{S}}(S, S')$ being the similarity of the inputs and $Y = \sigma_{\mathcal{R}}(\varphi(S), \varphi(S'))$ denoting the similarity of the associated outcomes, is then defined on the probability space $(\mathcal{S} \times \mathcal{S}, \mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}})$ as the mapping

$$(s, s') \mapsto \big(\sigma_{\mathcal{S}}(s, s'), \sigma_{\mathcal{R}}(\varphi(s), \varphi(s'))\big) .$$

Let $\mu_Z \overset{\mathrm{df}}{=} Z(\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}})$ be the induced probability measure on $D_{\mathcal{S}} \times D_{\mathcal{R}}$ and define $\mu_X$ on $D_{\mathcal{S}}$ and $\mu_Y$ on $D_{\mathcal{R}}$ in the same way. We shall use notations such as $(X = x)$ for events $X^{-1}(x)$ and $\mu_{Y|(X=x)} \overset{\mathrm{df}}{=} Y((\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}})(\cdot \,|\, X^{-1}(x)))$ to denote corresponding conditional probabilities. We also make use of intuitive notations such as $\mathbb{P}(Z = z)$ for $\mu_Z(z)$ or $\mathbb{P}(Y = y \,|\, X = x)$ for $\mu_{Y|(X=x)}(y)$. Besides, we employ the same symbol for a probability measure and the related distribution, i.e., we write $\mu_{\mathcal{S}}(s)$ rather than $\mu_{\mathcal{S}}(\{s\})$.

Remark 4.3. (1) The random variables $S$ and $R = \varphi(S)$ can be thought of as being defined over the same probability space as $Z$ under the mapping $(s, s') \mapsto s$ resp. $(s, s') \mapsto \varphi(s)$. We then obviously have $\mu_S = S(\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}}) = \mu_{\mathcal{S}}$ and $\mu_R = (\varphi \circ S)(\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}}) = \varphi(\mu_{\mathcal{S}})$.

(2) Subsequently, conditional probabilities which do actually not exist may appear in certain expressions (e.g., in a sum of measures). The probability

$$\mathbb{P}(Y = y \,|\, X = x, R = r),$$

for instance, is not well-defined unless there exist cases $s, s' \in \mathcal{S}$ with $\sigma_{\mathcal{S}}(s, s') = x$ and $\varphi(s) = r$ exist. Such probabilities and corresponding measures should simply be ignored. □

**Definition 4.4 (probabilistic similarity profile).** Consider a CBI setup $\Sigma$ and let $\mathcal{P}(D_{\mathcal{R}})$ denote the class of probability measures over $D_{\mathcal{R}}$. The mapping

$$H_\Sigma : D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{R}}) \,,\; x \mapsto \mu_{Y\,|\,(X=x)}$$

is called the probabilistic similarity profile (PSP) of $\Sigma$. □

The probabilistic similarity profile $H_\Sigma$ provides a much more precise picture of the similarity structure of a CBI setup $\Sigma$ than a (deterministic) similarity profile $h_\Sigma$ does. For each degree of similarity $x \in D_{\mathcal{S}}$, it specifies the probability distribution $\mu_{Y\,|\,(X=x)}$ of the similarity of outputs, i.e., of the random variable $Y$, given that the similarity of two inputs is $x$. Compared to this, the function $h_\Sigma$ provides only a lower bound to the support of $Y$:

$$h_\Sigma(x) = \inf \left\{ y \in D_{\mathcal{R}} \,|\, \mu_{Y\,|\,(X=x)}(y) > 0 \right\}.$$

**Definition 4.5 (stochastic dominance).** Let $\Omega \subset [0, 1]$ and denote by $\mathcal{P}(\Omega, \mathcal{A})$ the class of probability measures over the measurable space $(\Omega, \mathcal{A})$.[3] The decumulative distribution function of $\mu \in \mathcal{P}(\Omega, \mathcal{A})$ is defined as

$$G_\mu : [0, 1] \longrightarrow [0, 1] \,,\; x \mapsto \mu(\Omega \cap [x, 1]).$$

A probability measure $\mu$ is said to dominate a measure $\mu'$ stochastically, $\mu' \preceq \mu$, if $G_{\mu'} \leq G_\mu$. □

**Definition 4.6 (probabilistic similarity hypothesis).** A probabilistic similarity hypothesis is identified by a mapping $H : D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{R}})$. Let $\Sigma$ be a CBI setup with probabilistic similarity profile $H_\Sigma$. The hypothesis $H$ is admissible (with respect to $\Sigma$) if $H(x) \preceq H_\Sigma(x)$ for all $x \in D_{\mathcal{S}}$. $H$ is called a *strict* probabilistic hypothesis if

$$\forall\, x, x' \in D_{\mathcal{S}} \,:\, x < x' \Rightarrow H(x) \preceq H(x'). \tag{4.7}$$

A hypothesis $H'$ is called *stronger* than $H$ if $H(x) \preceq H'(x)$ for all $x \in D_{\mathcal{S}}$ and $H'(x_0) \not\preceq H(x_0)$ for at least one $x_0 \in D_{\mathcal{S}}$. □

[3] $\mathcal{P}(\Omega)$ stands for $\mathcal{P}(\Omega, 2^\Omega)$ if $\Omega$ is countable.

The stochastic dominance relation $\preceq$ over $\mathcal{P}(\Omega)$ is a natural generalization of the $\leq$-relation over $\Omega \subset [0,1]$. A hypothesis $H$ is admissible if it is "pessimistic" enough in the sense that it never over-estimates the probability that, for any $0 \leq \alpha \leq 1$, the similarity of the outcomes associated with two inputs is equal to or larger than $\alpha$. Within the probabilistic setting, the CBI hypothesis should be understood in the sense that "similar inputs *probably* have similar outcomes." Apparently, this is a special case of the non-deterministic version of the CBI hypothesis, suggesting that similar inputs are *likely* to have similar outcomes (cf. Chapter 1). A PSP gives a precise meaning to this assumption. In fact, it clarifies the meaning of "likely" in terms of probability distributions, and depicts its dependency on the similarity of inputs. Note that the strict version of a probabilistic hypothesis corresponds to the claim that, for all $0 \leq \alpha \leq 1$, the *more* similar two inputs are, the *larger* the probability will be that the associated outputs are at least $\alpha$-similar.

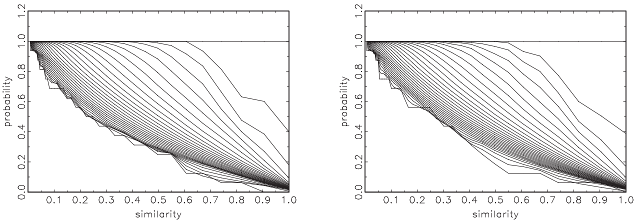


**Fig. 4.3.** Decumulative distribution functions associated with the probabilistic similarity profiles of the setup $\Sigma_1$ (left) and the setup $\Sigma_2$ (cf. Example 4.7).

Example 4.7. The probabilistic similarity profiles of the setups $\Sigma_1$ and $\Sigma_2$ introduced in Example 2.5 (cf. Section 2.5) are plotted in Fig. 4.3. More precisely, the pictures show the decumulative distribution functions $G_{H_{\Sigma_1}(x)}, G_{H_{\Sigma_2}(x)}$ for $x \in D_{\mathcal{S}}$. As can be seen, (4.7) with $H = H_{\Sigma_1}$ resp. $H = H_{\Sigma_2}$ holds true with a few exceptions,[4] i.e., the CBI hypothesis holds "almost" true in the strict sense. Fig. 4.4 depicts the decumulative distribution functions for the setups $\Sigma_1^*$ and $\Sigma_2^*$, for which an output corresponds to the (optimal) solution of an ILP. For these setups, the CBI hypothesis holds indeed true in the strict sense. □

Example 4.8. The probabilistic similarity profile of the CBI setup $\Sigma_3$, as defined in Example 4.1, is shown in Table 4.1. For $H_{\Sigma_3}$, the condition (4.7) is indeed completely satisfied, i.e., the CBI hypothesis applies in the strict sense. This can

[4] It should be taken into account that the number of observations is relatively small for some degrees of similarity.

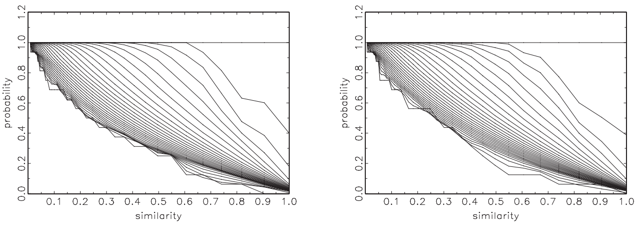


**Fig. 4.4.** Decumulative distribution functions associated with the probabilistic similarity profiles of the setup $\Sigma_1^*$ (left) and the setup $\Sigma_2^*$ (cf. Example 4.7).

| $x$ | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| $\mu_{Y \mid (X=x)}(0)$ | 0.11 | 0.08 | 0.07 | 0.05 | 0.04 | 0.03 | 0.02 | 0.00 |
| $\mu_{Y \mid (X=x)}(1/2)$ | 0.47 | 0.46 | 0.44 | 0.42 | 0.38 | 0.35 | 0.34 | 0.00 |
| $\mu_{Y \mid (X=x)}(1)$ | 0.42 | 0.47 | 0.50 | 0.53 | 0.58 | 0.62 | 0.64 | 1.00 |
| $G_{H_{\Sigma_3}(x)}(0)$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $G_{H_{\Sigma_3}(x)}(1/2)$ | 0.89 | 0.92 | 0.94 | 0.95 | 0.96 | 0.97 | 0.98 | 1.00 |
| $G_{H_{\Sigma_3}(x)}(1)$ | 0.42 | 0.47 | 0.50 | 0.53 | 0.58 | 0.62 | 0.65 | 1.00 |

**Table 4.1.** Probabilistic similarity profile of the setup $\Sigma_3$ and the corresponding decumulative distribution functions (cf. Example 4.8).

be gathered immediately from the decumulative distribution functions which are shown in the same table.[5] □

### 4.1.2 Generalized probabilistic profiles

In Chapter 3, we have motivated several alternative definitions of the concept of a similarity profile. We are now going to introduce corresponding probabilistic versions.

**Definition 4.9 ($(n,k)$-PSP).** Consider a CBI setup $\Sigma$ with associated memory $\mathcal{M} \sim (\mu_{\mathcal{S}\times\mathcal{R}})^n$, an input $S_0 \sim \mu_{\mathcal{S}}$, and $\langle S, \varphi(S)\rangle \sim \mu^{uni}_{\mathcal{N}^{ex}_k(\mathcal{M},S_0)}$, where $\mu^{uni}_{\mathcal{N}^{ex}_k(\mathcal{M},S_0)}$ denotes the uniform measure over the (extended) $k$-selection $\mathcal{N}^{ex}_k(\mathcal{M},S_0)$ (cf. Definition 3.9). Moreover, let the random variables $X$ and $Y$ be given by

$$X = \sigma_{\mathcal{S}}(S, S_0), \quad Y = \sigma_{\mathcal{R}}(\varphi(S), \varphi(S_0)).$$

The mapping

$$H^{(n,k)}_{\Sigma} : D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{S}}),\ x \mapsto \mu_{Y|(X=x)}$$

[5] Due to rounding errors not all columns in Table 4.1 sum up to 1.

is called the $(n,k)$-probabilistic similarity profile ($(n,k)$-PSP) of the setup $\Sigma$.[6] Thus, $[H_{\Sigma}^{(n,k)}(x)](y)$ is the probability that the similarity between $\varphi(S)$ and $\varphi(S_0)$ is $y$, given that the similarity between $S$ and $S_0$ is $x$, and that $S$ is chosen at random from the $k$ most similar cases in $\mathcal{M}$. □

The definition of an $(n,k)$-PSP is based on the same idea as the definition of an $(n,k)$-similarity profile (cf. Definition 3.9). Two inputs $S, S_0 \in \mathcal{S}$, the similarity relation between which is explored, are no longer chosen at random according to $\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}}$. Rather, $S$ is selected randomly from the set of inputs in the memory which are most similar to $S_0$. For the special case $k = 1$ the concept of an $(n,k)$-PSP corresponds to rules of the following form: "If the similarity between a new input and an input of maximum similarity is $x$, then the similarity between the corresponding outcomes is distributed according to $H_{\Sigma}^{(n,k)}(x)$."

**Definition 4.10 ($\mathcal{M}$-PSP).** Consider a setup $\Sigma$ with *fixed* memory $\mathcal{M}$ and let $S_0 \sim \mu_{\mathcal{S}}$, $\langle S, \varphi(S)\rangle \sim \mu_{\mathcal{M}}^{uni}$. Moreover, let $X = \sigma_{\mathcal{S}}(S, S_0)$ and $Y = \sigma_{\mathcal{R}}(\varphi(S), \varphi(S_0))$. The mapping

$$H_{\Sigma}^{\mathcal{M}} : D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{R}}),\ x \mapsto \mu_{Y|(X=x)}$$

is called the $\mathcal{M}$-probabilistic similarity profile ($\mathcal{M}$-PSP) of $\Sigma$. □

**Definition 4.11 (($\mathcal{M}, k$)-PSP).** Consider a setup $\Sigma$ with a *fixed* memory $\mathcal{M}$, an input $S_0 \sim \mu_{\mathcal{S}}$, and $\langle S, \varphi(S)\rangle \sim \mu_{\mathcal{N}_k^{ex}(\mathcal{M}, S_0)}^{uni}$. Moreover, let $X = \sigma_{\mathcal{S}}(S, S_0)$ and $Y = \sigma_{\mathcal{R}}(\varphi(S), \varphi(S_0))$. The mapping

$$H_{\Sigma}^{(n,k)} : D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{S}}),\ x \mapsto \mu_{Y|(X=x)}$$

is called the $(\mathcal{M}, k)$-probabilistic similarity profile ($(n,k)$-PSP) of $\Sigma$. □

**Definition 4.12 (local PSP).** Consider a setup $\Sigma$ and a *fixed* input $s \in \mathcal{S}$ and let $S_0$ be distributed according to $\mu_{\mathcal{S}}$. Moreover, let $X_s = \sigma_{\mathcal{S}}(s, S_0)$, $Y_s = \sigma_{\mathcal{R}}(\varphi(s), \varphi(S_0))$. The local probabilistic similarity profile associated with $s$, or $s$-PSP, is defined as

$$H_{\Sigma}^{s} : D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{R}}),\ x \mapsto \mu_{Y_s|(X_s=x)}.$$

A collection $H_{\Sigma}^{\mathcal{M}} = \{H_{\Sigma}^{s} \,|\, \langle s, \varphi(s)\rangle \in \mathcal{M}\}$ of local profiles is called a local $\mathcal{M}$-PSP. □

One verifies that the (global) PSP (cf. Definition 4.4) is a (pointwise) weighted average of the local profiles associated with individual cases:

[6] In fact, it may happen that some conditional measures do actually not exist. Then, the PSP of order $(n,k)$ is well-defined only on a subset of $D_{\mathcal{S}}$. The same remark applies to Definitions 4.10 and 4.11 below.

$$\forall\, x \in D_{\mathcal{S}} \,:\, H_{\Sigma}(x) \propto \sum_{s \in \mathcal{S}} \alpha(s,x) \cdot H_{\Sigma}^{s}(x), \tag{4.8}$$

where $H_{\Sigma}$ denotes the PSP of a setup $\Sigma$, and $H_{\Sigma}^{s}$ is the local PSP associated with $s \in \mathcal{S}$. Moreover, $\alpha(s,x) = \mu_{\mathcal{S}}(s) \cdot [X_s(\mu_{\mathcal{S}})](x)$ for all $s \in \mathcal{S}$, where $X_s : \mathcal{S} \longrightarrow D_{\mathcal{S}}$ denotes the mapping $s' \mapsto \sigma_{\mathcal{S}}(s,s')$.

## 4.2 Case-based inference, probabilistic reasoning, and statistical inference

Within the probabilistic setting of this chapter, the memory $\mathcal{M}$ as well as the information structures $\mathsf{SST}(\mathcal{M}, s_0)$ and $\mathsf{OST}(\mathcal{M}, s_0)$ appearing in (4.2) and (4.3) are random variables, the distribution of which can be derived from the measure (4.6). By combining the variables which constitute the similarity structure (cf. Definition 3.13) into one vector $Z_S$, we obtain a random variable defined over the probability space $(\mathcal{S}^{n+1}, (\mu_{\mathcal{S}})^{n+1})$. As before, we use the notation $\mu_{Z_S} \overset{\mathrm{df}}{=} Z_S((\mu_{\mathcal{S}})^{n+1})$. We denote by $Z_S - A$ the vector $Z_S$ reduced by a set $A$ of variables. For instance, $Z_S - \{X_{01}, X_{02}\}$ marks the vector $Z_S$ reduced by $X_{01}$ and $X_{02}$. Moreover, we denote by $Z_O$ the vector which combines the values associated with the outcome structure of a CBI problem.

The fact that the information used within the process of case-based inference can be seen as data emerging from a well-defined stochastic process allows for relating CBI to probabilistic reasoning. At the same time this framework makes the application of methods from *statistical inference* within the context of case-based reasoning possible. Let us illustrate these important aspects by means of two examples.

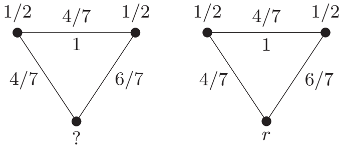


**Fig. 4.5.** Illustration of the outcome structures occurring in Example 4.13.

Example 4.13. For $n = 2$ the similarity structure corresponds to a vector $Z_S = (X_{01}, X_{02}, X_{12}, Y_{12})$, and the outcome structure defines the (extended) vector $Z_O = Z_S \cup (R_1, R_2)$. Suppose $\mathcal{M} = (\langle (3,14), 1/2\rangle, \langle (4,17), 1/2\rangle)$ and $s_0 = (5,17)$ in connection with the CBI setup $\Sigma_3$ defined in Example 4.1. Thus, $Z_S$ and $Z_O$ are realized by $z_S = (4/7, 6/7, 4/7, 1)$ and $z_O = z_S \cup (1/2, 1/2)$, respectively. For a certain value $r \in \mathcal{R}$ we then have

$$\mathbb{P}(R_0 = r \,|\, \mathsf{OST}(\mathcal{M}, s_0)) = \frac{(\mu_\mathcal{S} \otimes \mu_\mathcal{S} \otimes \mu_\mathcal{S})(\mathcal{S}_2)}{(\mu_\mathcal{S} \otimes \mu_\mathcal{S} \otimes \mu_\mathcal{S})(\mathcal{S}_1)}, \tag{4.9}$$

where $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{S} \times \mathcal{S} \times \mathcal{S}$ are the sets of triples of instances which are compatible (i.e., which can be "matched") with the first and the second structure in Fig. 4.5, respectively. This way, we obtain the following probabilities:

| $r$ | $0$ | $1/2$ | $1$ |
|---|---|---|---|
| $\mathbb{P}(R_0 = r \| Z_O = z_O)$ | 0.319 | 0.566 | 0.115 |

Observe that we have derived this result by means of a simple enumeration of the sets $\mathcal{S}_1$ and $\mathcal{S}_2$. Of course, this kind of inference is not in line with the idea of case-based reasoning. Besides, the information provided by a PSP does not permit the (exact) derivation of probabilities which condition on the complete outcome structure. As will be seen below, the probabilistic approach to CBI will generally result in approximations of these probabilities. □

Example 4.14. In order to illustrate the applicability of (inductive) statistical methods to CBI, let us consider two examples, namely the test of a simple hypothesis and a problem of parameter estimation related to the similarity structure of a CBI setup. In connection with our setup $\Sigma_3$, a hypothesis might be given, e.g., by the supposition that $\pi \geq 0.7$, where $\pi = \mathbb{P}(Y = 1 \,|\, X = 6/7)$ is the probability that two directly neighbored instances have identical outcomes. Based on a number of (independently) observed pairs of inputs, well-known statistical methods can be employed for realizing such a test procedure.

Recall that each value $H_\Sigma(x)$ of a PSP corresponds to a probability distribution on $\mathcal{R}$. In the case of $\Sigma_3$, such a value and, hence, the value $H(x)$ of a corresponding hypothesis $H$ can simply be specified as a stochastic vector

$$\big([H(x)](0), [H(x)](1/2), [H(x)](1)\big)\,.$$

If the set $\mathcal{R}$ of outputs is large or even infinite, however, it might be advantageous to specify $H(x)$ as a parameterized distribution, i.e., by means of a parameter vector which identifies a probability distribution. The specification of $H(x)$, i.e., the estimation of $H_\Sigma(x)$, then turns out to be a problem of parameter estimation. This example already suggests that the task of (probabilistic) case-based learning, i.e., the learning of a PSP, can be seen as a problem of statistical inference. We shall expand on this point in Section 4.3. □

Fig. 4.6 provides an overview of the probabilistic approach to CBI. Essentially, this approach realizes a process of *probabilistic reasoning in similarity space*[7] plus respective transformations between the *instance level* and the *similarity level*. The structure of this process corresponds to the one of constraint-based CBI (cf. Section 3.2):

[7] This contrasts with other probabilistic approaches to case-based inference [237, 275, 274, 369] which generally use a more implicit model of the CBI principle.

$$\begin{tikzcd}
(\mathcal{S}, \mathcal{R}, \varphi) \arrow[d, swap, "\begin{matrix}\sigma_{\mathcal{S}}\\ \sigma_{\mathcal{R}}\end{matrix}"] & (\mathcal{M}, s_0) \arrow[d, swap, "\begin{matrix}\sigma_{\mathcal{S}}\\ \sigma_{\mathcal{R}}\end{matrix}"] & \eta \\
H_{\Sigma} & Z_{\mathcal{S}} \arrow[l, swap, "\text{statistical inference}"] \arrow[r, swap, "\text{probabilistic reasoning}"] & \mathcal{C} \arrow[u, swap, "\sigma_{\mathcal{R}}^{(-1)}"]
\end{tikzcd}$$

**Fig. 4.6.** Illustration of the probabilistic approach to CBI which is a generalization of the constraint-based approach illustrated in Fig. 3.2.

- In a first step, the problem $\langle \Sigma, s_0 \rangle$ is again characterized at the *similarity level* by means of its *similarity structure*. In fact, the profile $H_{\Sigma}$ and the structure $Z_{\mathcal{S}}$ can be seen as the respective "image" of the system $(\mathcal{S}, \mathcal{R}, \varphi)$ and the (extended) memory $(\mathcal{M}, s_0)$ under the transformation defined by the similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$. This mapping realizes a projection from an often high-dimensional (and non-numeric) instance space $\mathcal{S} \times \mathcal{R}$ into the two-dimensional similarity space $D_{\mathcal{S}} \times D_{\mathcal{R}}$, which is usually more accessible to analytical methods. Still, this projection is not (information-)theoretically justified as is, say, principal components analysis in statistics are. Rather, it is guided by the heuristic assumption that the similarity structure of the problem $\langle \Sigma, s_0 \rangle$ represents useful information which is contained (implicitly) in the similarity measures.
- The similarity structure $Z_{\mathcal{S}}$ plays the role of *statistical data* within the CBI process. Moreover, the hypothesis $H$ defines the *stochastic model* which explains the occurrence of such structures and which underlies the reasoning process. In a second step, CBI makes use of this model and the given data in order to derive a (probabilistic) characterization of the unknown outcome $\varphi(s_0)$. This characterization, which corresponds to a probability measure $\mu \in \mathcal{P}(D_{\mathcal{R}})$ resp. a class $\mathcal{C}$ of such measures, is *implicit* in the sense that it is expressed in terms of similarity degrees, i.e., it does not refer to the output itself.
- Finally, the (probabilistic) information about similarity degrees has to be interpreted in the light of observed outcomes. That is, it has to be transformed into information at the *instance level*. In Fig. 4.6, this transformation is indicated by the (pseudo-)inverse mapping $\sigma_{\mathcal{R}}^{(-1)}$. It will be discussed in more detail in subsequent sections.

As already mentioned in Example 4.14, the task of *case-based learning*, if understood as the estimation of the similarity profile $H_{\Sigma}$, corresponds to *statistical inference*. Comparing Fig. 3.2 and Fig. 4.6 reveals that passing from constraint-based CBI to probabilistic CBI essentially means replacing a similarity profile $h_{\Sigma}$

by a probabilistic profile $H_{\Sigma}$, constraint-propagation by probabilistic reasoning, and sets (representing constraints) by uncertainty measures.

Realizing a transformation between the (high-dimensional) instance space and the similarity space does not mean that CBI leaves any information out of account: The information available at the instance level is utilized when computing degrees of similarity. In fact, passing from the complete description of instances to degrees of similarity is common to all inference schemes based on the NN principle and simply corresponds to an intermediate step of the complete inference procedure. It is a major assumption of this principle that similarity degrees represent the essential information in a condensed form. Observe, however, that our method does not only consider similarity degrees. In fact, it uses an additional concept, namely the similarity profile which provides an explicit model of the (otherwise implicitly used) NN principle. This way, CBI combines instance-based and model-based reasoning: Given a new query, each instance provides a prediction by making use of a model in the form of a similarity hypothesis.

The discussion so far has shown that the fields of case-based reasoning and probabilistic reasoning/statistical inference can benefit from each other. Firstly, a probabilistic interpretation of the CBI hypothesis allows for realizing CBI in the form of probabilistic reasoning and case-based learning as statistical inference. Secondly, Section 3.5 has shown that CBI can support statistical inference, or can even be interpreted as inductive statistical reasoning by itself.

More generally, a probabilistic formalization of case-based inference can be seen as a step toward an extended (probabilistic) approach to statistical reasoning. Classical statistical methods are principally based on the same kind of experience as case-based reasoning, namely a sequence of observations. As a decisive difference, however, let us mention that statistical methods concentrate on the probability distribution of the respective random variables directly and assume these random variables to be distributed identically. In other words, observations are generated under *identical* conditions. The occurrence of observations is explained by making assumptions about the data-generating process,[8] and conclusions about this process are drawn from the *frequency* of observed cases. As opposed to this, case-based reasoning assumes observations to be generated under conditions which are at most *similar*, i.e., it takes different data-generating processes into account.[9] Roughly speaking, it assumes only similarly instead of identically distributed random variables. This becomes apparent especially in connection with the non-deterministic framework proposed in Section 2.4.2, where an input determines the distribution of a random variable. That is, an *individual* probability distribution is associated with each input. A similarity hypothesis establishes a (probabilistic) relationship between these distributions, which in turn

[8] Consider *Bayesian cluster analysis*, where an individual data-generating process is associated with each cluster (= input), as an example. These processes, however, are completely independent of each other.

[9] One might argue that – stricto sensu – observations are never generated under identical conditions, at best under conditions which appear (approximately) identical from a certain perspective [299].

allows for making a connection between outcomes from different data-generating processes. Thus, case-based inference does not only make statistical assumptions about a data-generating processes itself, but also about the *relation* between several (similar) processes.[10] This way, it combines the concepts of frequency and similarity. Further points of contact between CBI and statistical inference which are in line with these ideas will be revealed in Section 4.3, where the (extended) problem of case-based learning is shown to provide an interesting approach to statistical modeling.

Let us conclude this section with an example illustrating the possibility of exploiting the similarity concept within an extended statistical framework. Consider two random samples $\mathcal{X}$ and $\mathcal{Y}$ induced by two data-generating processes. These shall be characterized by (unknown) parameters $\theta_1$ and $\theta_2$, respectively. Now, suppose the two processes to be similar in the sense that $\mathit{\Delta} = |\theta_1 - \theta_2|$ is small. This might suggest to use not only the sample $\mathcal{X}$ in order to estimate $\theta_1$, but also to include the sample $\mathcal{Y}$ (at least to some extent). Indeed, we can shown the following result.

**Proposition 4.15.** Let $\theta_1^e = f_1(\mathcal{X})$ and $\theta_2^e = f_2(\mathcal{Y})$ be unbiased estimators of two parameters $\theta_1$ resp. $\theta_2$, where $\mathbb{V}(\theta_1^e) > 0$. We then have $\mathsf{MSE}(\tilde{\theta}_1^e) < \mathsf{MSE}(\theta_1^e)$ for the estimator

$$\tilde{\theta}_1^e = \frac{\theta_1^e + \alpha\,\theta_2^e}{1+\alpha} \tag{4.10}$$

if the conditions $0 \leq \alpha \leq 1$ and

$$\alpha < \frac{2\,\mathbb{V}(\theta_1^e)}{\mathit{\Delta}^2 + \mathbb{V}(\theta_2^e) - \mathbb{V}(\theta_1^e)} \tag{4.11}$$

with $\mathit{\Delta} = |\theta_1 - \theta_2|$ are satisfied. □

**Proof.** Some transformations show that condition (4.11) is equivalent to

$$\frac{\alpha^2\,\mathit{\Delta}^2}{(1+\alpha)^2} + \frac{1}{(1+\alpha)^2}\,\left(\mathbb{V}(\theta_1^e) + \alpha^2\,\mathbb{V}(\theta_2^e)\right) < \mathbb{V}(\theta_1^e). \tag{4.12}$$

Note that $\mathsf{MSE}(\theta_1^e)$ is simply given by the variance of $\theta_1^e$ (i.e., by the right-hand side of (4.12)) since $\theta_1^e$ is assumed to be unbiased. Since

$$\begin{aligned}\mathsf{bias}(\tilde{\theta}_1^e) &= \left(\mathbb{E}(\tilde{\theta}_1^e) - \theta_1\right)^2 \\ &= \left(\frac{\theta_1 + \alpha\,\theta_2}{1+\alpha} - \theta_1\right)^2 = \frac{\alpha^2\,\mathit{\Delta}^2}{(1+\alpha)^2}.\end{aligned}$$

and

$$\mathbb{V}(\tilde{\theta}_1^e) = \frac{1}{(1+\alpha)^2}\,\left(\mathbb{V}(\theta_1^e) + \alpha^2\,\mathbb{V}(\theta_2^e)\right),$$

the left-hand side in (4.12) corresponds to $\mathsf{MSE}(\tilde{\theta}_1^e)$. □

[10] It is worth mentioning that assumptions of this kind have *implicitly* been made for a long time, e.g., when assuming that a family of probability density functions $f_\theta$ depends continuously on the parameter (= input) $\theta$ [73].

According to Proposition 4.15 the biased estimator (4.10) is superior to the unbiased estimator $\theta_1^e$ in the sense of the *mean squared error* criterion.[11] Roughly speaking, the increased sample size (of $|\mathcal{X}| + \alpha\,|\mathcal{Y}|$) leads to a reduced variance, and this compensates for the bias of $\tilde{\theta}_1^e$.

## 4.3 Learning probabilistic similarity hypotheses

In Section 4.2, we already pointed out that the learning of probabilistic similarity hypotheses can be seen as a statistical problem, namely that of estimating a probability distribution from observed data. In this section, we shall consider this problem in more detail. However, for the following reasons our treatment of the subject will remain superficial: Firstly, the estimation of probability distributions is by now a well-developed subfield of statistics, and a large number of methods is already available. Secondly, it would be difficult to develop standard techniques in connection with CBI, since different applications will generally call for different estimation methods.

### 4.3.1 Simple hypotheses and credible case-based inference

Despite its simplicity, the idea of approximating (deterministic) similarity profiles in terms of step functions, as proposed in Section 3.4, turned out to be useful and eventually produced the inference scheme of credible case-based inference. Now, this representation of similarity hypotheses can easily be extended to the probabilistic setting. Let $A_k$ be an interval in the representation (3.23) of hypotheses. Moreover, let $U_k$ be the set of similarity degrees $\sigma_{\mathcal{R}}(r_\imath, r_\jmath)$ such that $\sigma_{\mathcal{S}}(s_\imath, s_\jmath) \in A_k$. Rather than assigning to the coefficient $\beta_k$ the minimum of $U_k$, as in (3.25), we now define this bound by the $(1-p)$-quantile of $U_k$, where $p$ is a usually small value such as 0.05. As an empirical quantile, $\beta_k$ is hence an estimation of the corresponding true quantile of the distribution $y \mapsto \mathbb{P}(Y = y \,|\, X \in A_k)$. We call the step function $h^p$ given by $h^p(x) = \beta_k$ for $x \in A_k$, with $\beta_k$ as defined above, the *empirical p-profile.*

Now, suppose that we employ $h^p$ in order to derive a prediction

$$\widehat{\varphi}_{h^p,\mathcal{M}}(s_0) = \bigcap_{\imath=1}^{k} \mathcal{N}_{h^p(\sigma_{\mathcal{S}}(s_\imath,s_0))}(r_\imath), \tag{4.13}$$

where $s_1 \ldots s_k$ are the $k$ nearest neighbors of the query input $s_0$. What is the level of confidence of this prediction? Unfortunately, we do not have enough information to compute the probability of an incorrect prediction exactly. In fact, the

[11] This criterion goes back to Gauss and is currently the most popular one. Alternative criteria include the *mean absolute deviation* proposed by Laplace and the *measure of closeness* developed by Pitman.

individual predictions derived from the $k$ neighbors are obviously not independent in a stochastic sense. (We shall return to this problem in later sections.)

Still, by making a simplifying independence assumption, as it is often made in statistics, for example in connection with the well-known naïve Bayes classifier, one might justify assigning the confidence level $(1-p)^k$ to the prediction (4.13). Our practical experience has shown that this level still underestimates the true confidence level in most applications.

Of course, probabilistic estimations of the above type can be derived for different values $p_1 < p_2 < \ldots < p_\ell$. Thus, by using this probabilistic variant of credible case-based inference one obtains a nested sequence

$$\widehat{\varphi}_{h^{p_\ell},\mathcal{M}}(s_0) \subseteq \widehat{\varphi}_{h^{p_{\ell-1}},\mathcal{M}}(s_0) \subseteq \ldots \subseteq \widehat{\varphi}_{h^{p_1},\mathcal{M}}(s_0) \qquad (4.14)$$

of credible output sets with associated confidence levels. As an advantage of this kind of "stratified" prediction note that it differentiates between predicted outcomes better than a single credible output set does: The outputs in $\widehat{\varphi}_{h^{p_\ell},\mathcal{M}}(s_0)$ are the *most likely* ones, those in $\widehat{\varphi}_{h^{p_{\ell-1}},\mathcal{M}}(s_0) \setminus \widehat{\varphi}_{h^{p_\ell},\mathcal{M}}(s_0)$ are somewhat less likely, and so on.

### 4.3.2 Extended case-based learning

While the above approach is a direct extension of the credible case-based inference scheme of Section 3.4, we are now going to reconsider the problem of learning a PSP from a more general point of view. Actually, the latter corresponds to the problem of learning a class

$$\{\mu_{Y \mid (X=x)} \in \mathcal{P}(D_\mathcal{R}) \,|\, x \in D_\mathcal{S}\} \qquad (4.15)$$

of (conditional) probability measures or a joint measure $\mu_Z$ over $D_\mathcal{S} \times D_\mathcal{R}$ from which the measures (4.15) can be derived. Since $D_\mathcal{S}$ and $D_\mathcal{R}$ are countable, we only have to deal with discrete probability distributions.

Consider a memory $\mathcal{M}$ of $n$ cases $\langle s_k, r_k \rangle$ $(1 \leq k \leq n)$. This memory defines an independent *instance sample*, i.e., a sample of instances drawn at random from $\mathcal{S} \times \mathcal{R}$. The data underlying the estimation of the PSP is then given by the set of similarity relations

$$\mathcal{X} = \left\{ (x_{\imath\jmath}, y_{\imath\jmath}) \,|\, 1 \leq \imath \leq \jmath \leq n \right\}, \qquad (4.16)$$

where $(x_{\imath\jmath}, y_{\imath\jmath}) = (\sigma_\mathcal{S}(s_\imath, s_\jmath), \sigma_\mathcal{R}(r_\imath, r_\jmath))$.

Learning a PSP based on a sample (4.16) can be seen as a probabilistic counterpart to the basic CBL problem discussed in Section 3.4 (cf. Definition 3.27). Let us now consider the corresponding extended version of case-based learning, namely the learning of an adequate similarity measure $\sigma_\mathcal{S}$ and a hypothesis simultaneously. This problem is of special interest in a probabilistic setting. Besides, the simultaneous learning of similarity measures and probabilistic hypotheses provides an interesting approach to statistical modeling and inference.

Consider a parameterized class $\{H_\theta \,|\, \theta \in \Theta\}$ of probabilistic similarity profiles and let $\mu^\theta_{Y|(X=x)}$ denote the associated conditional probability measures, i.e.

$$\mu^\theta_{Y|(X=x)} = H_\theta(x).$$

Likewise, let $\{\sigma^\gamma_{\mathcal{S}} \,|\, \gamma \in \Gamma\}$ be a parameterized class of similarity measures. We assume the measure $\sigma_{\mathcal{R}}$ to be fixed. This assumption is necessary for technical reasons. Otherwise, the resulting model would have too many degrees of freedom, and a reasonable adaptation would not be possible. Besides, the assumption is not very critical since a reasonable definition of the similarity of outputs is possible for most applications. In fact, it is generally the specification of the similarity of inputs which is more difficult.

Let $\mathcal{X}$ be a sample (4.16) of similarity relations derived from a memory $\mathcal{M}$. More precisely, the elements of $\mathcal{X}$ are triples

$$(s, s', y) \in \mathcal{S} \times \mathcal{S} \times D_{\mathcal{R}},$$

where $y = \sigma_{\mathcal{R}}(\varphi(s), \varphi(s'))$. The similarity degrees $x = \sigma_{\mathcal{S}}(s, s')$ are still unknown, since the measure $\sigma_{\mathcal{S}}$ has not yet been defined. Now, consider the likelihood function $\lambda : \Gamma \times \Theta \longrightarrow \Re_{\geq 0}$ which specifies the probability of observing the similarity degrees $y$ given the respective inputs:

$$\begin{aligned}\lambda(\gamma, \theta) &= \prod_{(s,s',y)\in\mathcal{X}} \mathbb{P}(Y = y \,|\, \gamma, \theta, s, s') \\ &= \prod_{(s,s',y)\in\mathcal{X}} \mu^\theta_{Y|(X=\sigma^\gamma_{\mathcal{S}}(s,s'))}(y),\end{aligned}$$

where the random variable $Y$ denotes the similarity of outputs. The similarity measure $\sigma_{\mathcal{S}}$ and the PSP can be estimated by maximizing this likelihood function. That is, $\sigma_{\mathcal{S}}$ is estimated by $\sigma^{\gamma_{ML}}_{\mathcal{S}}$ and the PSP by $H_{\theta_{ML}}$, where $\gamma_{ML}$ and $\theta_{ML}$ denote the respective ML estimations.

The main difference between the case-based approach to statistical modeling outlined above and classical statistical methods is the structural assumption underlying the data-generating process (cf. Section 2.4). In case-based models, these assumptions are given in the form of the the CBI hypothesis. This hypothesis is used for explaining observations and thus plays a role somewhat similar to, say, the assumption of a linear relationship between input variables and output variables in regression analysis. A hypothesis related to the probabilistic similarity profile corresponds to the stochastic model. In linear regression, this model is specified by the linear structure and the distribution of an error term.

Thus, classical methods assume a (statistical) relationship between input variables and output variables. Typical models in economics or the social sciences, for instance, try to explain a certain attribute related to an individual by means of several other attributes of the same individual. As opposed to this, case-based models employ the attributes by more indirect means. Namely, these attributes specify the similarity relations between statistical entities, which in turn are utilized for explaining observations (cf. the discussion in Section 4.2).

Such models might hence be preferable if an observation is not really explained by the input variables. Let us consider an example from the field of economics, where the explaining (predictor) variables correspond to certain properties of a product and the variable to be predicted is the number of produced units. If it can be assumed that production is strongly influenced by the latest fashion, one will observe different production rates for the same product over time. That is, the number of produced units is actually not a function of the properties of the product. Of course, it is still possible to define such a function, but the related model would be valid only for a certain point of time. In other words, the corresponding statistical model does not describe a (time-)invariant relationship between variables. As opposed to this, the (rather plausible) CBI assumption, saying that similar products are produced in similar scope, remains valid over time. The CBI principle suggests not to guide the estimation of the current production rate of a product by its properties but by the current production rate of similar products.

Let us mention that the meaning of the parameters of a (parameterized) similarity measure is comparable, say, to that of the coefficients of a linear function in regression analysis. From an application-oriented point of view, the estimated measure might even be more interesting than the predictions themselves. Namely, this measure reveals the decisive properties which qualify statistical entities as being similar (with respect to the output variable). Consider again the above example and suppose that a skirt which is characterized as (long, tight, red) is found to be more similar to (long, tight, blue) than to (short, tight, red) as far as the number of produced units is concerned. This finding might suggest that the length of a skirt is a more important property than the color.

A more general application for which case-based inference seems appropriate is time series analysis. Consider a time series $(x(t))_{t \in T}$ and let inputs $s$ correspond to time points $t$, i.e., $\mathcal{S} = T$.[12] Moreover, let outputs be given by respective states $x(t)$. According to the general assumption underlying time series analysis, the state at time $t$ is determined by previous states and additional (external) influences which are modeled as a random variable. Thus, the output is actually not determined by the attributes of the input, i.e., the index of time. The CBI principle, however, generally holds true, at least if the influence of the random component is not too strong. Consider as an example a simple random walk with $T = \mathfrak{N}$ and $X(t+1) = X(t) + Z(t)$, where $(Z(t)+1) \sim \mathrm{BV}(2, 1/2)$. It is readily verified that

[12] Instead of data ordered by time one might also consider spatial data [316]. In that case $\mathcal{S}$ is of higher dimension.

$$\mathbb{P}(\Delta(x(t), x(t')) = \delta) = c(\delta) \binom{2|t-t'|}{\delta + |t-t'|} \left(\frac{1}{2}\right)^{2|t-t'|},$$

where the distance measure $\Delta$ is given by $(a, b) \mapsto |a - b|$, $c(\delta) = 1$ if $\delta = 0$ and $c(\delta) = 2$ otherwise. The PSP associated with these probabilities (and similarity measures related to the distance $| \cdot |$) shows that the CBI hypothesis applies extremely well.[13]

Let us finally mention that classical statistical models and case-based models can well be combined. Suppose, for instance, that a linear regression model is indeed appropriate, i.e., that a certain output variable can be explained by the linear combination of certain input variables. Moreover, consider a number of populations and suppose that each population gives rise to a different linear model, i.e., to a different vector of coefficients. Moreover, assume that the populations can be compared somehow, based on a set of further attributes. One might then consider the (CBI) assumption that similar populations have similar models. Note that the case-based model thus defined does no longer refer to the same input and output variables as the regression models. Rather, it is the class of regression models itself (i.e., the coefficients which identify such models) which constitute the set of outcomes of the case-based model. In this sense, the latter can be seen as a kind of meta-model characterizing the structure of a class of classical models. This way, it becomes possible to establish a link between estimation results of different populations, i.e., to support the estimation of one model based on the estimation results of other models or on observations related to these models.

## 4.4 Experiments with regression and label ranking

This section is meant to convey a first idea of how CBI can be applied to prediction problems and how it performs in practice. To this end, we present some experiments, in which we compared our approach to standard IBL (nearest neighbor estimation). It should be noted in advance, however, that a fair comparison is difficult, especially since the methods provide predictions of different kind. For example, the main purpose of CBI is to derive estimations in the form of *credible sets*, whereas IBL aims at producing good *point estimations* in the first place. As a consequence, standard IBL and CBI are not directly comparable. And indeed, the main purpose of our studies is not to show that one approach is better than the other one, but instead that CBI can reasonably complement standard IBL. Besides, the experiments are intended to support the theoretical results of the previous sections and to underpin our claim that CBI combines advantages from both instance-based and model-based learning.

We performed experiments for two types of prediction problems, namely regression (sections 4.4.1 and 4.4.2) and so-called label ranking (section 4.4.3). In the

[13] Observe that *causality* does not really matter in connection with CBI, in the sense that a value $x(t)$ can well be used for reasoning about $x(t')$ even if $t' < t$.

case of regression, a training example is a tuple $\langle s, r \rangle$, where $s = (s_1 \ldots s_m)$ is a vector of values for the input attributes, numerical or nominal, and $r$ is a value for the (numerical) output attribute. As a similarity measure, we used

$$\sigma_{\mathcal{S}}(x, y) \overset{\text{df}}{=} \exp\left( -\gamma \frac{1}{m} \sum_{\imath} d(x_\imath, y_\imath) \right), \tag{4.17}$$

where the distance $d(\cdot)$ is defined as $|x_\imath - y_\imath|$ for numerical attributes and assumes values 0 and 1 for ordinal features (i.e., $d(x_\imath, y_\imath) = 0$ if $x_\imath = y_\imath$ and $= 1$ otherwise). To guarantee that all attributes do approximately have the same influence – a point of critical importance in IBL [221] – each input attribute is first re-scaled linearly to the unit interval. To facilitate the interpretation of quality measures, we re-scaled the output attribute in the same way.

Since our main objective is to compare IBL and CBI under equal conditions, we refrained from "tuning" both methods. Particularly, we neither included feature selection nor feature weighting.[14] Besides, we did not put much effort in optimizing the constant $\gamma$ in (4.17); $\gamma = 5$ seemed to produce reasonable results, and we used this value throughout our experiments. The partition of the unit interval underlying the similarity hypothesis in CBI was always defined as a simple equi-width partition of size 10 for the global version and (since there are less training examples in the local approach) of size 5 for the local variant.

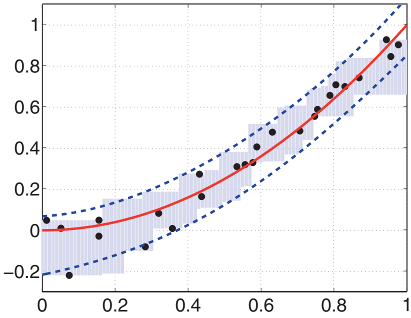


**Fig. 4.7.** Approximation of $x \mapsto x^2$ (solid line) in the form of a confidence band, using CBI (shaded region) and linear regression (region between dashed lines). The examples are indicated by black points.

### 4.4.1 Regression: artificial data

The first example is a simple regression problem and mainly serves an illustration purpose. The function to be learned is given by the polynomial $x \mapsto x^2$. Moreover,

[14] It is well-known that irrelevant features can badly deteriorate instance-based learning methods and, on the other hand, that feature weighting can greatly improve performance [396].

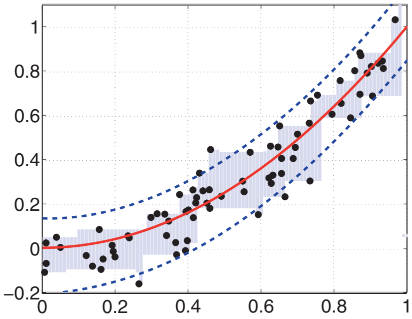


**Fig. 4.8.** Approximation of $x \mapsto x^2$ (solid line) in the form of a confidence band, using CBI with local profiles and linear regression (region between dashed lines).

$n$ training examples $\langle s_\imath, r_\imath \rangle$ are given, where the $s_\imath$ are uniformly distributed in $\mathcal{S} = [0, 1]$ and the associated outcomes $r_\imath$ are normally distributed with mean $(s_\imath)^2$ and standard deviation $1/10$. As mentioned above, we employed (4.17) with $\gamma = 5$ as a similarity measure for both inputs and outputs. Given a random sample (memory) $\mathcal{M}$, we first induce a similarity hypothesis for an underlying equi-width partition of size $m = 5$. A case-based approximation of the mapping $x \mapsto x^2$ is then derived from this hypothesis and the memory $\mathcal{M}$ (more precisely, a prediction $\widehat{\varphi}_{h,\mathcal{M}}(s)$ was derived for all $s \in \{0, 0.01, 0.02 \ldots 1\}$). Note that each prediction is simply an interval, so the case-based approximation (union of these intervals) yields a *confidence band* for the true mapping $x \mapsto x^2$. Fig. 4.7 shows a typical inference result for $n = 25$. Moreover, Fig. 4.8 shows a result for $n = 75$, using CBI with local similarity profiles.

According to our theoretical estimation, the degree of confidence for $n = 25$ is (at least) 16/26. This, however, is only a lower bound, and empirically (namely by averaging over 1,000 experiments) we found that the level of confidence is almost 0.9. To draw a comparison with standard statistical techniques, the figures also show the 0.9-confidence band obtained for the regression estimation (and the same samples). As can be seen, CBI yields predictions of roughly the same precision, and CBI with local profiles is even slightly more precise. This finding was also confirmed for estimation problems with other functions and input spaces of higher dimension.

In this connection, it should again be mentioned that linear resp. polynomial regression makes much more assumptions than CBI. Especially, the type of function to be estimated must be specified in advance: Knowing that this function is a polynomial of degree 2 in our example, we took the model $x \mapsto \beta_0 + \beta_1 x + \beta_2 x^2$ as a point of departure and estimated the coefficients $\beta_\imath$. Usually, however, such knowledge will not be available. For instance, the performance of LR becomes much worse due to typical overfitting effects when adapting a polynomial of degree $k > 3$ to the data. Moreover, the confidence band for LR is only valid if the error terms follow a normal distribution (as they do in our case but not in general).

### 4.4.2 Regression: real-world data

We also applied CBI to several real-world data sets from the UCI repository and the Statlib archive.[15] The data is summarized in table 4.2.

| | name | size | # var. |
|---|---|---|---|
| 01 | breast-tumor | 277 | 1/8 |
| 02 | cholesterol | 297 | 6/7 |
| 03 | cleveland | 297 | 6/7 |
| 04 | cpu | 209 | 6/1 |
| 05 | housing | 506 | 12/1 |
| 07 | pharynx | 193 | 1/10 |
| 08 | sensory | 576 | 0/11 |
| 09 | strike | 625 | 5/1 |
| 10 | bodyfat | 252 | 14/0 |
| 11 | pollution | 60 | 15/0 |
| 12 | pw-linear | 200 | 10/0 |
| 13 | auto-price | 159 | 15/0 |
| 15 | bolts | 40 | 7/0 |
| 16 | cloud | 108 | 4/2 |
| 18 | fruitfly | 125 | 2/2 |
| 19 | lowbwt | 189 | 7/2 |
| 20 | fishcatch | 71 | 5/2 |
| 21 | echo-months | 61 | 6/3 |
| 22 | quake | 2178 | 3/0 |
| 23 | auto-mpg | 392 | 4/0 |

**Table 4.2.** Data sets used in the experiments: name, number of examples, number of predictor variables (numerical/nominal).

In order to test the effectiveness of the probabilistic version of CBI, we have applied this approach to the data sets with different values for the parameter $p$ (namely $p = 0, 0.02, 0.04$). The following performance measures were derived by means of a leave-one-out cross-validation:

1. The *correctness* or empirical confidence (CONF) measured in terms of the relative frequency of correct predictions (predicted interval covers true value).
2. The *precision* of predictions (PREC) measured in terms of the average length of a predicted interval.
3. The *mean absolute error* (MAE) measured in terms of the average distance between the true value and the point estimation (center of the interval).

As a neighborhood size for CBI we used $k = 20$. Again, note that this parameter is less important in CBI than in $k$-NN estimation. As mentioned previously, dissimilar neighbors will often hardly influence the prediction in terms of a credible

[15] `http://www.ics.uci.edu/~mlearn`, `http://lib.stat.cmu.edu/`

set. And indeed, we observed that even though varying this parameter has an effect for small $k$, increasing $k$ beyond $\approx 15$ hardly changed the results.

The results for this series of experiments are summarized in table 4.3. As can be seen, the use of probabilistic bounds yields an extreme gain of precision at the cost of a mostly slight deterioration of the confidence. This finding, which basically holds true for all data sets, clearly provides strong evidence for the effectiveness of the probabilistic extension of CBI: By varying the parameter $p$, a smooth tradeoff between confidence and precision can be achieved. Regarding the quality of the CBI point estimation, the influence of $p$ is less strong, though in general, more precise estimations come along with a slightly more accurate point estimation.

Admittedly, there are some data sets for which CBI performs poorly, either in terms of confidence or in terms of precision or both. Looking at the characteristics of these data sets, there are two plausible explanations. Firstly, confidence and precision is weak if the size of the data set is too small. Of course, this is natural, since statistically confident and precise predictions cannot be made on the basis of sparse data. Secondly, CBI seems to have problems with data sets in which nominal attributes prevail. As a plausible explanation, note that in this case there exist only a small number of different similarity degrees $\sigma_{\mathcal{S}}(x, y)$. If these degrees are not well distributed over the unit interval, an equi-with partition is likely to produce a poor and unbalanced similarity profile. In this case, the use of an *adaptive* partition (in line with equi-frequency histograms) seems to be advised, an option that we did not exploit so far but that should definitely be given a try.

| | CONF | PREC | MAE | CONF | PREC | MAE | CONF | PREC | MAE |
|---|---|---|---|---|---|---|---|---|---|
| 01 | 0.9856 | .8817 | .1680 | .8159 | .5879 | .1696 | .7329 | .4041 | .1714 |
| 02 | 0.9663 | .5918 | .1271 | .8013 | .3310 | .0909 | .6599 | .2398 | .0909 |
| 03 | 1.0000 | .8695 | .3136 | .9024 | .5800 | .2555 | .7576 | .3981 | .2344 |
| 04 | 0.9809 | .0665 | .0187 | .8278 | .0404 | .0177 | .7608 | .0274 | .0187 |
| 05 | 1.0000 | .6134 | .0904 | .8538 | .3185 | .0689 | .7787 | .2146 | .0643 |
| 07 | 0.9896 | .7895 | .2069 | .8083 | .5596 | .1807 | .7202 | .4670 | .1791 |
| 08 | 1.0000 | .9184 | .1194 | .8333 | .4118 | .1250 | .7326 | .2826 | .1222 |
| 09 | 0.9888 | .7758 | .3506 | .8368 | .1551 | .0727 | .7296 | .1104 | .0608 |
| 10 | 0.9802 | .3946 | .0607 | .8333 | .2095 | .0620 | .6984 | .1881 | .0663 |
| 11 | 0.9500 | .4974 | .1173 | .7500 | .3267 | .1232 | .6333 | .2682 | .1140 |
| 12 | 0.9800 | .5526 | .0955 | .8200 | .3267 | .0908 | .7300 | .2727 | .0921 |
| 13 | 0.9623 | .2484 | .0583 | .7547 | .1404 | .0509 | .6792 | .1223 | .0544 |
| 15 | 0.9250 | .5802 | .2021 | .6250 | .3903 | .1801 | .4750 | .1965 | .1483 |
| 16 | 0.9537 | .2956 | .0714 | .7963 | .2157 | .0739 | .7315 | .1848 | .0779 |
| 18 | 0.9520 | .8692 | .2411 | .7760 | .5743 | .2004 | .6240 | .4549 | .1947 |
| 19 | 0.9577 | .5292 | .0934 | .7937 | .3281 | .0950 | .6508 | .2584 | .0983 |
| 20 | 0.9437 | .2014 | .0506 | .8732 | .1876 | .0544 | .7183 | .1330 | .0487 |
| 21 | 0.9344 | .7245 | .2520 | .8033 | .6601 | .2542 | .7049 | .5443 | .2322 |
| 23 | 0.9923 | .6389 | .1180 | .8316 | .3775 | .0956 | .7015 | .2855 | .0905 |

**Table 4.3.** Results for probabilistic CBI: Confidence, precision, and mean absolute error of predictions for $p = 0$ (left), $p = 0.02$ (middle), and $p = 0.04$ (right).

| | 1-NN | 3-NN | 5-NN | 7-NN | 9-NN |
|---|---|---|---|---|---|
| 01 | .4200 | .4626 | .4195 | .4009 | .3841 |
| 02 | .4093 | .4641 | .4561 | .4675 | .4544 |
| 03 | .6664 | .6097 | .5906 | .6330 | .6302 |
| 04 | .6386 | .6719 | .6448 | .6882 | .6663 |
| 05 | .2555 | .4667 | .4782 | .4976 | .5532 |
| 07 | .4554 | .5771 | .5881 | .5964 | .6172 |
| 08 | .6387 | .6549 | .6310 | .6136 | .6338 |
| 09 | .5718 | .6936 | .7131 | .7202 | .7153 |
| 10 | .1626 | .2128 | .2394 | .2200 | .2303 |
| 11 | .4279 | .1371 | .2812 | .3102 | .3739 |
| 12 | .1154 | .3805 | .5039 | .4627 | .4535 |
| 13 | .5752 | .7048 | .6967 | .6977 | .7094 |
| 15 | .7270 | .7618 | .7600 | .6863 | .6584 |
| 16 | .7001 | .7730 | .7332 | .7509 | .7337 |
| 18 | .4511 | .5618 | .5145 | .4881 | .4462 |
| 19 | .2180 | .3312 | .3412 | .3601 | .3513 |
| 20 | .4904 | .4205 | .3090 | .3384 | .2839 |
| 21 | .4783 | .5938 | .6030 | .6412 | .6642 |
| 22 | .3050 | .3441 | .3497 | .3478 | .3558 |
| 23 | .4407 | .4522 | .4713 | .4466 | .4586 |

**Table 4.4.** Statistical correlation between precision of CBI predictions and mean absolute error of the $k$-NN estimations for $k = 1, 3, 5, 7, 9$.

We also found that the CBI point estimations are on average slightly inferior to the point estimations produced by standard $k$-NN estimation (see also table 4.5 below), even though there are some exceptions where the former are even better than the latter. Nevertheless, table 4.4 shows the statistical (Pearson) correlation between the precision (PREC) of CBI estimations and the mean absolute error of the standard $k$-NN estimations. As can be seen, there is a strong positive correlation between these two quantities throughout. This finding suggests that the width of the CBI confidence interval is a good indicator of the accuracy of a $k$-NN prediction. Consequently, it might be an interesting idea to complement the latter by the former, i.e., to take the $k$-NN prediction as a point estimation and the CBI prediction as a confidence interval.

In a second series of experiments, we have employed the local version of CBI. The results are summarized in table 4.5. As it was to be expected from our theoretical analysis, predictions become more precise but less confident in comparison with the global version of CBI. Apart from that, it is interesting to note that local CBI yields extremely good point estimations. In fact, more often than not, these point estimations are better than those of standard $k$-NN. Recalling that CBI is actually not intended to produce point estimations, at least not in the first place, this is a surprisingly good an indeed unexpected result.

### 4.4.3 Label ranking

In principle, CBI can be applied to classification problems in the same way as to regression problems. In this connection, however, it should be mentioned that CBI is useful only if the number of class labels is not too small, since otherwise

| | CONF | PREC | MAE | 1-NN | 3-NN | 5-NN | 7-NN | 9-NN |
|---|---|---|---|---|---|---|---|---|
| 01 | 0.8159 | .3700 | .1387 | .2349 | .1798 | .1747 | .1736 | .1738 |
| 02 | 0.7407 | .2159 | .0782 | .1263 | .1020 | .0960 | .0926 | .0899 |
| 03 | 1.0000 | .2104 | .0934 | .1692 | .1546 | .1549 | .1556 | .1531 |
| 04 | 0.8038 | .0530 | .0202 | .0165 | .0241 | .0288 | .0306 | .0309 |
| 05 | 0.8261 | .1416 | .0516 | .0683 | .0573 | .0612 | .0653 | .0680 |
| 07 | 0.7254 | .2926 | .1135 | .1989 | .1578 | .1429 | .1394 | .1367 |
| 08 | 0.8281 | .1326 | .0663 | .1431 | .1279 | .1288 | .1236 | .1178 |
| 09 | 0.8432 | .0913 | .0328 | .0344 | .0279 | .0296 | .0291 | .0293 |
| 10 | 0.8254 | .1608 | .0580 | .0686 | .0523 | .0533 | .0556 | .0568 |
| 11 | 0.8333 | .2263 | .0865 | .1234 | .1166 | .1044 | .1098 | .1087 |
| 12 | 0.8600 | .1899 | .0696 | .1138 | .0879 | .0832 | .0823 | .0851 |
| 13 | 0.7547 | .1203 | .0424 | .0498 | .0502 | .0539 | .0562 | .0589 |
| 15 | 0.7000 | .1718 | .0754 | .1482 | .1285 | .1112 | .1311 | .1397 |
| 16 | 0.7407 | .2432 | .0916 | .0887 | .0805 | .0773 | .0872 | .0963 |
| 18 | 0.7680 | .4226 | .1458 | .2317 | .1690 | .1641 | .1597 | .1580 |
| 19 | 0.8201 | .2087 | .0724 | .1074 | .0961 | .0901 | .0889 | .0891 |
| 20 | 0.7042 | .0928 | .0342 | .0305 | .0396 | .0583 | .0634 | .0639 |
| 21 | 0.7541 | .4917 | .1660 | .1999 | .2020 | .1965 | .1869 | .1878 |
| 22 | 0.9752 | .4235 | .1472 | .1710 | .1448 | .1412 | .1402 | .1396 |
| 23 | 0.8571 | .2757 | .0727 | .0900 | .0765 | .0750 | .0735 | .0742 |

**Table 4.5.** Results for CBI with local profiles: Confidence, precision, and mean absolute error of predictions; mean absolute error for $k$-NN point estimations with $k = 1, 3, 5, 7, 9$.

the prediction of credible (label) sets does hardly make sense. Anyway, CBI somehow unifies diverse types of prediction problems. Moreover, as it does not make strong structural assumptions for the output space $\mathcal{R}$ but only requires a similarity measure $\sigma_{\mathcal{R}}$ to be given, it is widely applicable and especially interesting for learning problems involving structured output spaces [372]. This section is meant to illustrate this point by applying CBI to the problem of *label ranking*.

The problem of label ranking has been introduced quite recently [184, 156] and can be considered as a generalization of standard classification. In the classification setting, each instance $s \in \mathcal{S}$ is associated with a single label $y \in \mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2 \ldots y_\ell\}$ is a finite set of class labels. Given a set of examples in the form of labelled instances $(s, y)$, the problem is to induce a classification function, i.e., a $\mathcal{S} \longrightarrow \mathcal{Y}$ mapping from the input to the output space. In label ranking, each instance $s$ is instead associated with a complete *ranking* (total order) of the labels $\mathcal{Y}$. Correspondingly, the problem is to learn a *ranking function* that maps instances to rankings over $\mathcal{Y}$.

A ranking can be expressed in terms of a permutation $\tau_s$ of $\{1, 2 \ldots \ell\}$, where $\tau_s(\imath) = \jmath$ if the class label $y_\imath$ has position $\jmath$ in the ranking associated with instance $s$. Thus, the output space $\mathcal{R}$ in our CBI framework can now be defined by the set of all permutations of $\{1, 2 \ldots \ell\}$. As a similarity measure $\sigma_{\mathcal{R}}$ we employ the well-known *Spearman rank correlation*:

$$\sigma_{\mathcal{R}}(\tau, \tau') \overset{\mathrm{df}}{=} 1 - \frac{6 \sum_{\imath=1}^{\ell} (\, \tau'(\imath) - \tau(\imath)\, )^2)}{\ell(\ell^2 - 1)}.$$

More specifically, since the rank correlation (just like the standard Pearson correlation for numerical attributes) assumes values in $[-1, 1]$, we use an affine transformation of this measure to the unit interval.

Since benchmark data for label ranking is not yet available, we generated such data from standard classification data in the following way: We first trained a naïve Bayes classifier[16] on a given classification data set. Then, for each example instance $s_\imath$, all the labels were ordered with respect to decreasing predicted class probabilities (in the case of ties, labels with lower indices are ranked first). By substituting the single labels contained in the original (multiclass) data set with the complete rankings, we finally obtained a label ranking data set as desired. As an aside, note that the problem of learning a ranking function may thus also be viewed as learning a qualitative replication of the naïve Bayes predictions.

In the following, we present results for the `glass` data, again a well-known benchmark from the UCI repository. This data set contains 9 predictive attributes. Since all these attributes are numerical, we again used (4.17) with $\gamma = 5$ as a similarity measure $\sigma_{\mathcal{S}}$. The number of classes in this data set is $\ell = 6$.[17] After the data has been transformed into a label ranking data set as described above, the same performance measures as in section 4.4.2 were derived, again by means of a leave-one-out cross-validation: The confidence (CONF) is again measured in terms of the relative frequency of correct predictions, i.e., predictions covering the true ranking. The precision (PREC) was measured in terms of the mean of $|\widehat{\varphi}_{h,\mathcal{M}}(s_0)| \cdot |\mathcal{R}|^{-1} = |\widehat{\varphi}_{h,\mathcal{M}}(s_0)| \cdot (\ell!)^{-1}$, i.e., the average relative size of the credible output set. Instead of the mean absolute error as derived in section 4.4.2, we determined the *mean similarity* (MS) between the true ranking and the point estimation of CBI (generalized median (3.31) of $\widehat{\varphi}(s_0)$).

| $k$ | $p$ | CONF | PREC | MS |
|---|---|---|---|---|
| 3 | .00 | 1.00 | 0.412 | 0.962 |
| 3 | .02 | 0.95 | 0.136 | 0.971 |
| 3 | .04 | 0.93 | 0.110 | 0.972 |
| 7 | .00 | 1.00 | 0.405 | 0.960 |
| 7 | .02 | 0.94 | 0.131 | 0.970 |
| 7 | .04 | 0.91 | 0.107 | 0.972 |
| 15 | .00 | 1.00 | 0.401 | 0.959 |
| 15 | .02 | 0.92 | 0.128 | 0.967 |
| 15 | .04 | 0.88 | 0.105 | 0.970 |

**Table 4.6.** Experimental results obtained by applying the probabilistic variant of CBI (cf. Section 4.3.1) to the label ranking version of the `glass` data.

Table 4.6 shows results for different sizes $k$ of the neighborhood, using the probabilistic version of CBI as an inference method. The results are quite comparable to those of Section 4.4.2. Again, the use of probabilistic bounds yields a considerable

[16] We employed the implementation of the Weka machine learning package [400].

[17] Actually, the number is 7, but since the fourth attribute never occurs we only considered the remaining 6.

gain of precision at the cost of a slight deterioration of the precision. The quality of the point estimations is not affected very much, but it seems that the more precise the credible output set, the better the point estimation derived thereof.

Table 4.7 summarizes the performance of standard IBL ($k$-nearest neighbor estimation) in terms of the mean similarity between predicted and true rankings. Here, IBL predictions were derived by the median of the query's $k$ nearest neighbors.[18] Again, it turns out that, regarding point estimations, CBI is more than competitive with standard IBL.

| $k$ | 1 | 3 | 5 | 7 | 15 |
|---|---|---|---|---|---|
| MS | 0.969 | 0.965 | 0.957 | 0.947 | 0.926 |

**Table 4.7.** Mean similarity between predicted and true rankings for standard IBL ($k$-nearest neighbor estimation).

## 4.5 Case-based inference as evidential reasoning

The symbol $\eta$ in Fig. 4.6, characterizing the prediction derived by means of probabilistic CBI, designates a normalized uncertainty measure (fuzzy measure [388]) over $\mathcal{R}$, i.e., a mapping $2^{\mathcal{R}} \mapsto [0,1]$ such that

- $\eta(\emptyset) = 0$, $\eta(\mathcal{R}) = 1$,
- $\forall\, A, B \in \mathcal{R} \, : \, A \subset B \Rightarrow \eta(A) \leq \eta(B)$.

In fact, the probabilistic approach to CBI will generally not allow for the derivation of a unique probability distribution on $\mathcal{R}$. This is caused by the two properties of CBI mentioned in Section 3.2.1. Firstly, the *indirect* derivation of predictions necessitates the *transformation* of constraints on similarity degrees into constraints on outcomes. Secondly, the *locality* of inference rules calls for the *combination* of probabilistic evidence obtained from individual cases. Both, the transformation as well as the combination of probabilistic constraints are part of the (pseudo-)inverse $\sigma_{\mathcal{R}}^{(-1)}$.

In the constraint-based setting of Chapter 3, evidence concerning similarity degrees is given in the form of lower similarity bounds, i.e., intervals of similarity degrees, and the transformation of this evidence is realized by means of the set-valued mapping (3.4). Moreover, the intersection of corresponding constraints on the output level is accomplished by a simple intersection. The derivation of a nested sequence (4.14) of credible output sets in Section 4.3.1 can be considered

[18] This is sometimes called the *set median* [215]. While the generalized median of a set $U \subset X$ of objects is an element of $X$, the set median is searched among the given objects only (the $k$ nearest neighbors in our case) and, hence, is an element of $U$.

as a direct probabilistic generalization of this approach, justified by a simplifying assumption of independence.

In this section, the problem of combining probabilistic evidence in connection with CBI will be considered in a more general context, namely as a *parallel combination of information sources*. The problem of combining concurrent pieces of (uncertain) evidence arises in many fields, such as robotics (sensor fusion) or knowledge-based systems (expert opinion pooling), and it has been dealt with in a probabilistic setting [153, 165] as well as alternative uncertainty frameworks [25, 44, 120, 207]. The combination of evidence derived from individual cases is perhaps best compared to that of expert opinion pooling. That is, each (observed) case is seen as an expert, and its prediction of the unknown outcome of the new input is interpreted as an expert statement. The task is to synthesize these statements.

A general framework for the parallel combination of information sources which seems suitable for our purpose has been introduced in [164]. A basic concept within this framework is that of an *imperfect specification*: Let $\Omega$ denote a set of alternatives consisting of all possible states of an object under consideration and let $\omega_0 \in \Omega$ be the actual (but unknown) state.[19] An imperfect specification of $\omega_0$ is a tuple $\Gamma = (\gamma, p_C)$, where $C$ is a (finite) set of *specification contexts*, $\gamma$ is a $C \longrightarrow 2^{\Omega}$ mapping, and $p_C$ is a probability measure over $C$.[20] The problem of combining evidence is then defined as generating one imperfect specification $\Gamma$ of $\omega_0$ from $n$ imperfect specifications $\Gamma_1, \ldots, \Gamma_n$, issued by $n$ different information sources.

From a semantical point of view, a specification context $c \in C$ can be seen as a physical or observation-related frame condition, and $\gamma(c)$ is the (most specific) characterization of $\omega_0$ that can be provided by the information source in the context $c$. Consider the testing of the equality of two numbers, realized in the form of a predicate $P(\alpha, \beta) \equiv (\alpha = \beta)$, as a simple example and let $\Omega = \{0,1\} \times \{0,1\}$. We can then distinguish the contexts $c_1$ and $c_2$ in which the predicate $P$ is true and in which $P$ is false, respectively, when being applied to $\omega_0$. This leads to $\gamma(c_1) = \{(0,0),(1,1)\}$ and $\gamma(c_2) = \{(0,1),(1,0)\}$.

The value $p_C(c)$ can be interpreted as an (objective or subjective) probability of selecting $c$ as a true context. An imperfect specification is thus able to model *imprecision* as well as *uncertainty*. The consideration of (probabilistic) uncertainty is accomplished by the probability measure $p_C$. Moreover, the modeling of imprecision is possible due to the fact that $\gamma$ is a *set-valued* mapping.

[19] The fact that $\omega_0$ can always be represented as an element of $\Omega$ is a consequence of the (often implicitly made) *closed world assumption* [348].

[20] Formally, an imperfect specification is nothing but a set-valued mapping on a probability space, a well-known concept in connection with random sets [82, 280, 360].

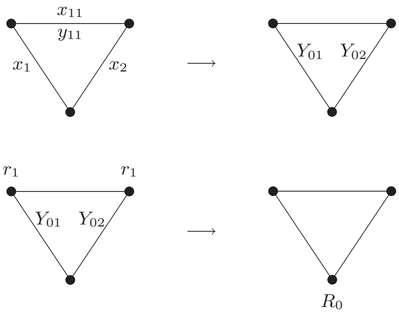


**Fig. 4.9.** Illustration of probabilistic CBI as a procedure consisting of two steps.

### 4.5.1 Transformation of probabilistic evidence

According to the indirect approach realized by CBI, evidence concerning outcomes is derived in two stages, where the second step consists of translating (probabilistic) evidence about similarity degrees into constraints on outcomes (cf. Fig. 4.9).

Within the probabilistic setting of this chapter, evidence concerning similarity degrees is given in the form of probability measures. Consider a probability measure $\mu$ over $D_{\mathcal{R}}$ which has been derived from a case $\langle s, r\rangle$ and which is taken as evidence about the similarity between $r$ and the unknown outcome $\varphi(s_0)$. When interpreting this case as an information source $\mathit{\Gamma} = (\gamma, p_C)$, the set of specification contexts is given by the set of possible degrees of similarity $x = \sigma_{\mathcal{S}}(s, s_0)$.[21] That is,

$$\begin{array}{lcl} C & = & D_{\mathcal{R}}, \\ \gamma(c) & = & \sigma_{\mathcal{R}}^{(-1)}(r, c), \\ p_C(c) & = & \mu(c), \end{array}$$

where

$$\sigma_{\mathcal{R}}^{(-1)}(r, c) \stackrel{\mathrm{df}}{=} \{r' \in \mathcal{R} \,|\, \sigma_{\mathcal{R}}(r, r') = c\} \tag{4.18}$$

for all $c \in C$. The set $\gamma(c)$ is obviously the most specific restriction of $\varphi(s_0)$ that can be derived in the context $c$, i.e., from the assumption that $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)) = c$ and the fact that $\varphi(s) = r$.

Let $\mathit{\Gamma} = (\gamma, p_C)$ denote the imperfect specification of an unknown outcome $\varphi(s_0)$ associated with a case $\langle s, \varphi(s)\rangle$. It may happen that $\gamma(c) = \emptyset$ for some $c \in C$, which means that $c$ cannot be a true context and that $\mathit{\Gamma}$ is contradictory [164]. It is then necessary to replace $\mathit{\Gamma}$ by a revised specification $\mathit{\Gamma}' = (\gamma', p_{C'})$. The latter is defined by

[21] For the sake of simplicity, we assume in this section that $D_{\mathcal{R}}$ is finite .

$$\begin{array}{rcl} C' & = & \{c \in C \,|\, \gamma(c) \neq \emptyset\}, \\ \gamma'(c') & = & \gamma(c'), \\ p_{C'}(c') & = & k \cdot p_C(c') \end{array}$$

for all $c' \in C'$, with $k$ being the normalization factor, i.e.

$$1/k = \sum_{c \in C \,:\, \gamma(c) \neq \emptyset} p_C(c). \tag{4.19}$$

Subsequently, the imperfect specification associated with a case $\langle s, r \rangle$ will always refer to the already revised specification.[22]

Note that the imperfect specification $\Gamma$ thus defined is closely related to the concept of a *mass distribution* in the belief function setting [336, 350]: Let $\mathsf{m} : 2^{\Omega} \longrightarrow [0,1]$ be a mass distribution over a set $\Omega$, i.e., $\mathsf{m}(\emptyset) = 0$ and $\sum_{A \subset \Omega} \mathsf{m}(A) = 1$. Moreover, let $\mathcal{A} = \{A_1, \ldots, A_n\} = \{A \subset \Omega \,|\, \mathsf{m}(A) > 0\}$ denote the (finite) set of *focal elements*. We can then associate an imperfect specification $\Gamma = (\gamma, p_C)$ with $\mathsf{m}$:

$$\begin{array}{rcl} C & = & \{c_1, \ldots, c_n\}, \\ \gamma(c_k) & = & A_k, \\ p_C(c_k) & = & \mathsf{m}(A_k) \end{array}$$

for all $1 \leq k \leq n$. The other way round, each imperfect specification $\Gamma = (\gamma, p_C)$ induces an (information-compressed[23]) representation in the form of a mass distribution $\mathsf{m}$, where

$$\mathsf{m}(A) = \sum_{c \in C \,:\, \gamma(c) = A} p_C(c) \tag{4.20}$$

for all $A \subset \Omega$ and $\mathsf{m}(A) > 0$ for a finite number of sets.

By making use of the relation between the mass function (4.20) and the imperfect specification $(\gamma, p_C)$ associated with a case $\langle s, \varphi(s) \rangle$, the evidence about the outcome $\varphi(s_0)$ derived from $\langle s, \varphi(s) \rangle$ can be represented in the form of a belief function $\mathsf{Bel}$ resp. an associated plausibility function $\mathsf{Pl}$ over $\mathcal{R}$, where

$$\mathsf{Bel}(A) = \sum_{B \subset A} \mathsf{m}(B), \qquad \mathsf{Pl}(A) = \sum_{B \cap A \neq \emptyset} \mathsf{m}(B)$$

for all $A \subset \mathcal{R}$. $\mathsf{Bel}(A)$ and $\mathsf{Pl}(A)$ define degrees of belief and plausibility that $\varphi(s_0)$ is an element of $A$, respectively. These values can also be interpreted as lower and upper probabilities. Since the imperfect specification and, hence, the mass distribution associated with $\langle s, \varphi(s) \rangle$ is derived from the outcome $\varphi(s)$ and the probability measure $H(\sigma_{\mathcal{S}}(s, s_0))$, the above belief (plausibility) function corresponds to a transformation $\sigma_{\mathcal{R}}^{(-1)}$ which is now a $\mathcal{R} \times \mathcal{P}(D_{\mathcal{R}}) \longrightarrow \mathcal{F}(\mathcal{R})$ mapping, where $\mathcal{F}(\mathcal{R})$ denotes, say, the class of normalized uncertainty measures over $\mathcal{R}$:

[22] We disregard cases for which $k$ in (4.19) is not well-defined.

[23] A mass function does not define a unique imperfect specification.

$$\mathsf{Bel} = \mathsf{Bel}(H, s_0) = \sigma_{\mathcal{R}}^{(-1)} \left(\varphi(s), H(\sigma_{\mathcal{S}}(s, s_0))\right). \tag{4.21}$$

This transformation defines a generalization of $\sigma_{\mathcal{R}}^{(-1)}$ in (4.18).

Remark 4.16. The transformation (4.18) defines a kind of equivalence (indistinguishability) relation: An information source $\langle s, r\rangle$ does not distinguish between outcomes $r'$ and $r''$ such that $\sigma_{\mathcal{R}}(r, r') = \sigma_{\mathcal{R}}(r, r'')$. Thus, the belief function (4.21) has a special structure. In fact, the focal elements define a partition of $\mathcal{R}$. □

Let $\Gamma = (\gamma, p_C)$ be the imperfect specification induced by a case $\langle s, \varphi(s)\rangle$. The application of a *generalized insufficient reason principle* [350] makes it possible to characterize $\varphi(s_0)$ by means of a probability measure $\mathbb{P}$ over $\mathcal{R}$. The latter is defined by

$$\mathbb{P}(A) \overset{\text{df}}{=} \sum_{c \in C \,:\, \gamma(c) \cap A \neq \emptyset} p_C(c) \cdot \frac{|A \cap \gamma(c)|}{|\gamma(c)|} \tag{4.22}$$

for all $A \subset \mathcal{R}$, where $|X|$ denotes the cardinality of the set $X$. This measure is also called *betting function*, a term referring to the use of (4.22) in the context of decision making [350].

### 4.5.2 Inference from individual cases

Suppose a new input $s_0 \in \mathcal{S}$ to be given and let $\langle s, \varphi(s)\rangle$ be an observed case (chosen at random according to $\mu_{\mathcal{S}\times\mathcal{R}}$). Translating probabilistic evidence referring to degrees of similarity into evidence about outcomes, as outlined in Section 4.5.1, leads to a prediction in the form of the belief function

$$\mathsf{Bel} = \sigma_{\mathcal{R}}^{(-1)} \left(\varphi(s), H\left(\sigma_{\mathcal{S}}(s, s_0)\right)\right) \tag{4.23}$$

over the set $\mathcal{R}$ of outputs. As already mentioned above, the transformation $\sigma_{\mathcal{R}}^{(-1)}$ realizes a $\mathcal{R} \times \mathcal{P}(D_{\mathcal{R}}) \longrightarrow \mathcal{F}(\mathcal{R})$ mapping and defines a generalization of (4.18). If CBI proceeds from a local $\mathcal{M}$-hypothesis, (4.23) becomes

$$\mathsf{Bel} = \sigma_{\mathcal{R}}^{(-1)} \left(\varphi(s), H^s\left(\sigma_{\mathcal{S}}(s, s_0)\right)\right). \tag{4.24}$$

As (4.23) and (4.24) show, the framework introduced in this section gives rise to the (uncertain) specification of $\varphi(s_0)$ in the form of a belief function over $\mathcal{R}$.

In Section 4.5.3, we shall discuss the problem of combining several predictions (4.23) which have been derived from individual cases of the memory $\mathcal{M}$. Though let us mention that one might also think of selecting merely one previous case (maximally similar to the new input) from $\mathcal{M}$ for solving a new problem. This strategy, which is common practice in CBR, obviously avoids any kind of combination problem. Observe, however, that inputs are no longer determined by means of pure random choice. In fact, the idea of drawing inferences from the most similar case is supported by the concepts of an $(n,k)$-PSP and an $(\mathcal{M},k)$-PSP which have been introduced in Section 4.1. Indeed, what is basically needed is the kind of inference rules provided by these similarity profiles for the special case $k=1$. Even though these inference rules appear more complicated (compare the formulation on page 138) the specification of a hypothesis related to an $(n,1)$-PSP or an $(\mathcal{M},1)$-PSP does not seem to be more demanding than the specification of a hypothesis related to a PSP (all the more if these hypotheses are derived by means of machine learning methods).

Consider a memory $\mathcal{M}$, a new input $s_0 \in \mathcal{S}$, and let $H$ be a hypothesis related to an $(n,1)$-PSP or an $(\mathcal{M},1)$-PSP. We can then derive the following counterpart to (4.23) (for the mapping $\mathcal{N}_1$ see Definition 3.8):

$$\mathsf{Bel} = \sigma_{\mathcal{R}}^{(-1)}\left(\varphi(\mathcal{N}_1(\mathcal{M},s_0)), H\left(\sigma_{\mathcal{S}}(s_0,\mathcal{N}_1(\mathcal{M},s_0))\right)\right).$$

| $x$ | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| $\mathbb{P}(Y=0 \mid X=x)$ | 0.14 | 0.10 | 0.03 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
| $\mathbb{P}(Y=1/2 \mid X=x)$ | 0.55 | 0.57 | 0.53 | 0.45 | 0.37 | 0.24 | 0.19 | 0.00 |
| $\mathbb{P}(Y=1 \mid X=x)$ | 0.31 | 0.33 | 0.44 | 0.54 | 0.62 | 0.75 | 0.81 | 1.00 |

**Table 4.8.** $(\mathcal{M},1)$-PSP for the setup $\Sigma_3$ and a memory $\mathcal{M}$ of size 10 (cf. Example 4.17).

Example 4.17. Table 4.8 shows the $(\mathcal{M},1)$-PSP $H_{\Sigma_3}^{(\mathcal{M},1)}$ for the setup $\Sigma_3$ and a memory $\mathcal{M}$ of size 10 which has been generated at random. Consider the following inference scheme: Given a new input $s_0$, the memory $\mathcal{M}$ and the profile $H_{\Sigma_3}^{(\mathcal{M},1)}$ are used in connection with (4.23) for deriving evidence about the unknown outcome $\varphi(s_0)$ in the form of a belief function. An estimation of $\varphi(s_0)$ is then chosen at random according to the betting function (4.22) associated with this belief function. It can be shown that this procedure yields a correct estimation with a probability of approximately 1/2. Thus, exploiting the memory $\mathcal{M}$ (the size of which is only 10) increases the probability of a correct classification from 1/3 (which corresponds to a random choice) to 1/2. The percentage of correct classifications is even larger when making a deterministic choice by simply selecting the class with the highest degree of plausibility. □

### 4.5.3 Combining evidence from several cases

After having discussed the transformation of probabilistic evidence and its utilization for deriving inference results from an individual case, let us now turn to the problem of combining evidence from several cases. That is, suppose we are given $n$ imperfect specifications of the unknown outcome $\varphi(s_0)$, which have been derived from a memory $\mathcal{M}$ containing $n$ cases $\langle s_1,\varphi(s_1)\rangle,\ldots,\langle s_n,\varphi(s_n)\rangle$ in connection with a probabilistic similarity hypothesis $H$. The task shall be to aggregate these (uncertain) pieces of evidence.

**The problem of interdependence.** Within the framework of Section 3, evidence derived from an individual case $\langle s, \varphi(s) \rangle$ is given in the form of the $\alpha$-neighborhood $\mathcal{N}_\alpha(\varphi(s))$, where $\alpha = h(\sigma_{\mathcal{S}}(s, s_0))$. This corresponds to a particular imperfect specification $\Gamma = (\gamma, p_C)$, namely

$$\begin{aligned} C &= D_{\mathcal{R}}, \\ \gamma(c) &= \mathcal{N}_c(\varphi(s)), \\ p_C(c) &= \begin{cases} 1 & \text{if} \quad c = h(\sigma_{\mathcal{S}}(s, s_0)) \\ 0 & \text{if} \quad c \neq h(\sigma_{\mathcal{S}}(s, s_0)) \end{cases} . \end{aligned}$$

The probability measure $p_C$ thus defined reveals the cautious character of the inference scheme (3.2). In fact, the complete probability mass is attached to the (most pessimistic) context $c = h(\sigma_{\mathcal{S}}(s, s_0))$, thus suggesting that the similarity between $s$ and $s_0$ might not be larger than the lower bound $h(\sigma_{\mathcal{S}}(s, s_0))$. Within the setting of this section, where $p_C$ is not restricted to one-point measures, one could think of generalizing the *conjunctive* combination (3.2) by considering the prediction of $r_0$ as a random set.

Suppose the similarity between $\varphi(s_0)$ and $\varphi(s_k)$ to be given by $y_k$, i.e.

$$\forall\, 1 \leq k \leq n \, : \, \sigma_{\mathcal{R}}(\varphi(s_0), \varphi(s_k)) = y_k. \tag{4.25}$$

We can then derive the prediction $\varphi(s_0) \in \widehat{\varphi}_{y,\mathcal{M}}(s_0)$, where $y = (y_1, \ldots, y_n)$ and

$$\widehat{\varphi}_{y,\mathcal{M}}(s_0) \overset{\text{df}}{=} \bigcap_{1 \leq k \leq n} \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), y_k). \tag{4.26}$$

This corresponds to a conjunctive combination of the individual (set-valued) predictions $\sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), y_k)$. Within our probabilistic setting, the vector of similarity degrees is actually a random variable $Y = (Y_1, \ldots, Y_n)$, and the related prediction (4.26) can hence be seen as a random set $\widehat{\varphi}_{Y,\mathcal{M}}(s_0)$.

This approach comes down to considering the $n$ cases as one information source, inducing the imperfect specification $\Gamma = (\gamma, p_C)$, where

$$\begin{aligned} C &= (D_{\mathcal{R}})^n, \\ p_C &= \mu(c), \\ \gamma(c) &= \textstyle\bigcap_{1 \leq k \leq n} \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), c_k) \end{aligned} \tag{4.27}$$

for all $c = (c_1, \ldots, c_n) \in C$. The measure $\mu$ is the joint probability over $(D_{\mathcal{R}})^n$ characterizing the occurrence of similarity vectors. That is, $\mu(y)$ is the probability of the event (4.25), where $y = (y_1, \ldots, y_n)$.

Example 4.18. Let us reconsider Example 4.13. The following table shows the probabilities $\mathbb{P}((Y_{01}, Y_{02}) = (y_{01}, y_{02}) \,|\, Z_S = z_S)$:

| | 0 | 1/2 | 1 |
|---|---|---|---|
| 0 | $\frac{238}{24203}$ | 0 | 0 |
| 1/2 | 0 | $\frac{6559}{24203}$ | 0 |
| 1 | 0 | 0 | $\frac{17406}{24203}$ |

By making use of this probability distribution and the observed outcomes $r_1 = r_2 = 1/2$, we obtain the imperfect specification $\Gamma = (\gamma, p_C)$, where $C = \{0, 1/2, 1\} \times \{0, 1/2, 1\}$, $p_C$ is specified by the above probabilities, and

$$\begin{aligned} \gamma((0,0)) &= \emptyset \\ \gamma((1/2,1/2)) &= \{0,1\} \\ \gamma((1,1)) &= \{1/2\}. \end{aligned}$$

Observe that $c = (c_1, c_2)$ is only a possible context if $c_1 = c_2$, i.e., values $\gamma((c_1, c_2))$ for which $c_1 \neq c_2$ are not relevant. Moreover, this imperfect specification has to be revised, since $\gamma((0,0)) = \emptyset$. The revised specification leads to a mass distribution $\mathsf{m}$ such that $m(\{1/2\}) = 0.726$ and $m(\{0,1\}) = 0.274$. For the induced belief function we have

$$\mathsf{Bel}(\{0\}) = \mathsf{Bel}(\{1\}) = 0.274, \quad \mathsf{Bel}(\{1/2\}) = 0.726.$$

As already mentioned before, these values can be seen as upper probabilities of the respective outcomes $r \in \mathcal{R}$. □

Treating $n$ cases as one information source in the sense of (4.27) is an obvious way of combining evidence. What makes things difficult, however, is the fact that the joint probability measure $\mu$ over $(D_{\mathcal{R}})^n$ and, hence, the probability $p_C$ in (4.27) are generally not known. It is also not possible to derive this measure from the information provided by a PSP. The PSP informs about the (conditional) distributions of *individual* similarity degrees, i.e., it specifies the (unknown) similarity $y_k$ between $\varphi(s_0)$ and $\varphi(s_k)$ by means of a probability measure, given the similarity of the respective inputs: $Y_k \sim \mu_{Y|(X=\sigma_{\mathcal{S}}(s_0,s_k))}$. It is by no means obvious, however, how to derive the joint measure $\mu_{Y|Z_S}$ which takes the information provided by the complete similarity structure into account. In fact, the random variables $X_{\imath\jmath}, Y_{\imath\jmath}, X_{0\jmath}$ $(1 \leq \imath < \jmath \leq n)$, which constitute the similarity structure, are not stochastically independent. For instance, the similarities $X_{\imath\jmath}$ $(1 \leq \imath < \jmath \leq n)$ between the inputs in the memory depict important information about dependency structures which cannot be taken from a PSP. Needless to say, extending the PSP to a probabilistic model which provides the required information is generally intractable due to the enormous number of joint measures $\mu_{Y|Z_S}$ which would have to be specified.

The aforementioned type of interdependence is already revealed by Example 4.13. In fact, one obtains completely different sets of "matching" (triples of) cases $\mathcal{S}_1$ and $\mathcal{S}_2$ in (4.9) when ignoring the similarity $x_{12}$ between the observed cases. However, the problem can also be grasped intuitively. Consider the two prediction

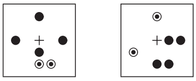


**Fig. 4.10.** Illustration of two prediction tasks.

tasks illustrated in Fig. 4.10. The left and the right picture in this figure show a partial neighborhood of the instance $s_0^1 = (8, 14)$ and the instance $s_0^2 = (24, 3)$, respectively. Interestingly enough, one obtains the same individual predictions in both cases and, therefore, the same overall prediction when combining them independently. Intuitively, however, the plausibility of $\varphi(s_0^1) = 1$ seems to be larger than the plausibility of $\varphi(s_0^2) = 1$.

This estimation is of course suggested by the locations of the neighbored points ($s_0^1$ is "surrounded" by black points).[24] In a more general context, where cases are not necessarily points in a Euclidean space, it is just the similarity among the observed cases which has an important influence on their individual contribution to the overall prediction. For example, the increase in evidence due to the observation of a second case, pointing in the same direction as a first case, does also depend on the similarity between these two cases. In fact, the more similar the two inputs are, the less astonishing it is that their outcomes are similar. This follows immediately from the CBI hypothesis itself. Thus, interdependence of the above type should not be ignored when taking this hypothesis for granted (as instance-based reasoning methods actually do).

As can be seen, it might not be advisable to consider individual cases as pieces of evidence which are *distinct* in the sense of [336, 349]. Therefore, we shall not apply the orthogonal sum operation as proposed by Dempster in order to combine individual predictions.

Example 4.19. Consider again Example 4.13. From the PSP in Table 4.1 we can take the measures $\mu_{Y|(X=x)} \in \mathcal{P}(D_\mathcal{R})$, i.e., the following (conditional) distributions:

| $y$ | 0 | 1/2 | 1 |
|---|---|---|---|
| $\mathbb{P}(Y_{01} = y \mid X_{01} = 4/7)$ | 0.04 | 0.38 | 0.58 |
| $\mathbb{P}(Y_{02} = y \mid X_{02} = 6/7)$ | 0.02 | 0.33 | 0.65 |

There is obviously now way of deriving the joint distribution of $Y_{01}$ and $Y_{02}$ (tabulated in Example 4.18) from these individual distributions without taking further information into account. The fact that the conditional probability of $(Y_{01}, Y_{02}) = (y_{01}, y_{02})$ is 0 whenever $y_{01} \neq y_{02}$, for instance, becomes obvious only from $y_{11} = 1$ and the special structure of $(\mathcal{R}, \sigma_\mathcal{R})$. □

[24] The example becomes especially convincing when thinking of black and white points as places with cloudy and sunny sky, respectively.

Remark 4.20. If the conditional probability measures

$$\mu_{Y|(Z_S=z_S)} \in \mathcal{P}((D_{\mathcal{R}})^n) \tag{4.28}$$

were known for all similarity structures $z_S$, the combination of evidence from different cases could be avoided completely. Since each measure (4.28) can be seen as a collection of inference rules, this would require some kind of "higher order CBI hypothesis." For instance, rules of the following form would have to be specified for each possible vector $z_s = (x_{01}, x_{02}, x_{11}, y_{11}) \in (D_{\mathcal{S}})^3 \times D_{\mathcal{R}}$ if $n = 2$: "Given two $x_{11}$-similar inputs with $y_{11}$-similar outcomes such that the similarity between the first resp. second input and the new input is $x_{01}$ resp. $x_{02}$, the similarities $Y = (Y_{01}, Y_{02})$ between the corresponding pairs of outcomes are distributed according to $\mu_{Y|(Z_S=z_s)}$." Even though this approach appears rather complex, we will take it up again in Section 4.7. □

**Convex combination of evidence.** If knowledge about the dependency structure is incomplete, a still reasonable way of combining evidence is to define the aggregated imperfect specification as the *convex combination* of the individual imperfect specifications [164].[25] Let $\Gamma_k = (\gamma_k, p_{C_k})$ denote the imperfect specification associated with the case $\langle s_k, \varphi(s_k)\rangle$, where

$$\begin{aligned} C_k &= \{k\} \times D_{\mathcal{R}}, \\ \gamma_k(c) &= \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), y), \\ p_{C_k}(c) &= \mu_{Y|(X=\sigma_{\mathcal{S}}(s_0,s_k))}(y) \end{aligned}$$

for all $c = (k, y) \in C_k$. Thus, an element $c = (k, y)$ specifies the context in which the $k$-th case is considered, and the similarity between the corresponding outcome $\varphi(s_k)$ and the unknown outcome $r_0$ is given by $y$. The convex combination $\Gamma = (\gamma, p_C)$ of $\Gamma_1, \ldots, \Gamma_n$ is then defined by

$$\begin{aligned} C &= C_1 \cup \ldots \cup C_n, \\ \gamma(c) &= \gamma_k(c), \\ p_C(c) &= \alpha_k \cdot p_{C_k}(c) \end{aligned} \tag{4.29}$$

for all $c = (k, y) \in C$, where $\alpha_k \geq 0$ $(1 \leq k \leq n)$ and $\alpha_1 + \ldots + \alpha_n = 1$.

Notice that the set of specification contexts in (4.29) is given by the *union* of the individual contexts, whereas it is defined as the *product* in (4.27). In fact, the incomplete specification (4.29) does not consider combined events (4.25) since the probabilities of these events are unknown. Rather, an interpretation of the convex combination $\Gamma$ can be given in terms of the following semantic model: Since a well-justified combination of the information sources cannot be accomplished,

[25] Often, other combination modes will actually be more appropriate, e.g., conjunctive or disjunctive pooling methods. In general, however, such methods assume more knowledge about the dependency structure or reliability of the information sources.

only one expert (case) is singled out and the corresponding belief (estimation) is adopted. The selection of the expert is realized by means of a random choice, where $\alpha_k$ is the probability of selecting the $k$-th expert. The convex combination $\Gamma$ can then be seen as the "expected belief" (i.e., the belief before the choice has been made). The weights $\alpha_k$ might be interpreted as an estimation of the relative reliability or specificity of the information sources. The question of how to determine these weights will be discussed in Section 4.6.

Now, consider a CBI problem $\langle \Sigma, s_0 \rangle$. Let $\mathsf{m}_k(H, s_0)$ and $\mathsf{Bel}_k(H, s_0)$ denote, respectively, the mass distribution and belief function induced by the $k$-th case $\langle s_k, \varphi(s_k) \rangle$, which corresponds to the $k$-th specification $\Gamma_k$. That is,

$$\mathsf{Bel}_k(H, s_0) = \sigma_{\mathcal{R}}^{(-1)} \left( \varphi(s_k), H(\sigma_{\mathcal{S}}(s_0, s_k)) \right), \tag{4.30}$$

with $H$ being a hypothesis related to $H_\Sigma$. The corresponding functions $\mathsf{m}$ and $\mathsf{Bel}$ associated with the convex combination (4.29) are then given by

$$\begin{aligned} \mathsf{m}(H, \mathcal{M}, s_0) &= \alpha_1 \cdot \mathsf{m}_1(H, s_0) + \ldots + \alpha_n \cdot \mathsf{m}_n(H, s_0), & (4.31) \\ \mathsf{Bel}(H, \mathcal{M}, s_0) &= \alpha_1 \cdot \mathsf{Bel}_1(H, s_0) + \ldots + \alpha_n \cdot \mathsf{Bel}_n(H, s_0). & (4.32) \end{aligned}$$

In plain words, combining evidence at the instance level comes down to deriving the convex combination of the belief functions induced by individual cases, where the weight of a case depends on characteristics such as similarity, typicality, or precision. Observe that the global hypothesis $H$ in (4.32) is replaced by the local hypotheses associated with the respective cases if CBI proceeds from a local $\mathcal{M}$-hypotheses $H^{\mathcal{M}}$:

$$\mathsf{Bel}(H^{\mathcal{M}}, \mathcal{M}, s_0) = \sum_{k=1}^{n} \alpha_k \cdot \mathsf{Bel}_k(H^{s_k}, s_0) \tag{4.33}$$

Given a setup $\Sigma$ with memory $\mathcal{M}$, a prediction (4.32) can principally be derived for all inputs in $\mathcal{S}$. This way, the case-based inference scheme can be generalized to a "belief function-valued" approximation of $\varphi$:[26]

$$\widehat{\varphi}_{H,\mathcal{M}} : \mathcal{S} \longrightarrow \mathcal{F}(\mathcal{R}) \, , \; s \mapsto \mathsf{Bel}(H, \mathcal{M}, s). \tag{4.34}$$

Of course, it is not necessary to derive a prediction (4.32) for those inputs which have already been observed and are stored in $\mathcal{M}$ since the corresponding outcome can simply be retrieved from the memory. That is, $\widehat{\varphi}_{H,\mathcal{M}}$ should actually be defined as

$$\widehat{\varphi}_{H,\mathcal{M}}(s) = \begin{cases} \mathsf{Bel}_{\{\varphi(s)\}} & \text{if} \quad \langle s, \varphi(s) \rangle \in \mathcal{M} \\ \mathsf{Bel}(H, \mathcal{M}, s) & \text{if} \quad \langle s, \varphi(s) \rangle \notin \mathcal{M} \end{cases}, \tag{4.35}$$

where $\mathsf{Bel}_{\{\varphi(s)\}}$ is the belief function focused on $\varphi(s)$: $\mathsf{Bel}_{\{\varphi(s)\}}(A) = 1$ for $A \supseteq \varphi(s)$ and $\mathsf{Bel}_{\{\varphi(s)\}}(A) = 0$ otherwise.

[26] This mapping corresponds to what is called an *extensional concept description* in instance-based learning [11].

## 4.6 Assessment of cases

A problem of practical relevance is the choice of the weights $\alpha_k$ in (4.31) and (4.32). As already mentioned above, these weights should reflect the reliability or quality of the individual information sources. At the same time, the determination of weights makes it possible to take the problem of interdependence (cf. Section 4.5.3) into account. This section is meant to suggest concrete criteria for defining the weights. Due to the involved nature of the problem, however, our results will not go beyond some heuristic approaches.

### 4.6.1 Similarity-weighted approximation

In locally weighted approximation [17] and weighted $k$NN [327], the influence of an observation is usually determined as a function of its distance to the query point. Thus, it is assumed that more similar cases are also more reliable or relevant in the current context. This idea suggests to define weights as normalized degrees of similarity:

$$\alpha_k = \sigma_{\mathcal{S}}(s_0, s_k) \left( \sum_{\imath=1}^{n} \sigma_{\mathcal{S}}(s_0, s_\imath) \right)^{-1} \tag{4.36}$$

if $\sum_{\imath=1}^{n} \sigma_{\mathcal{S}}(s_0, s_\imath) > 0$ and $a_k = 1/n$ otherwise.

Remark 4.21. Note that (4.36) might appear questionable as soon as we have $\sigma_{\mathcal{S}}(s_0, s_k) = 1$ for some $1 \leq k \leq n$. However, when assuming that $\sigma_{\mathcal{S}}$ is *separating* in the sense that $(\sigma_{\mathcal{S}}(s, s') = 1) \Leftrightarrow (s = s')$ this means that $s_0$ itself has already been observed. Therefore, the output $\varphi(s_0)$ is retrieved from the memory according to (4.35), i.e., the determination of weights is actually not necessary. If, on the other hand, $\sigma_{\mathcal{S}}$ is not separating (and $s_0$ is not stored in $\mathcal{M}$), it can indeed happen that two completely similar inputs have different outcomes. In this case (4.36) does again make sense. □

| | $r = 0$ | $r = 1/2$ | $r = 1$ |
|---|---|---|---|
| $[\mathsf{PI}(H_\Sigma, \mathcal{M}, (5,4))](\{r\})$ | 0.61 | 0.36 | 0.03 |
| $[\mathsf{PI}(H_\Sigma, \mathcal{M}, (14,15))](\{r\})$ | 0.31 | 0.39 | 0.30 |
| $[\mathsf{PI}(H_\Sigma^{\mathcal{M}}, \mathcal{M}, (5,4))](\{r\})$ | 0.91 | 0.07 | 0.02 |
| $[\mathsf{PI}(H_\Sigma^{\mathcal{M}}, \mathcal{M}, (14,15))](\{r\})$ | 0.37 | 0.49 | 0.14 |

**Table 4.9.** Prediction based on a memory of size $n = 25$ (cf. Example 4.22).

Example 4.22. In connection with Example 4.1, we have derived the prediction (4.32) with $H = H_\Sigma$ for two new inputs, namely $s_0^1 = (5,4)$ and $s_0^2 = (14,15)$. To this end, we have chosen a memory $\mathcal{M}$ of size $n = 25$ at random. Table 4.9

shows values of the associated plausibility functions. As can be seen, the highest degree of plausibility is assigned to the true outcomes $\varphi(s_0^1) = 0$ and $\varphi(s_0^2) = 1/2$, respectively. By making use of the local $\mathcal{M}$-PSP $H_{\Sigma}^{\mathcal{M}}$, we have also derived the prediction (4.33). The corresponding degrees of plausibility are again shown in Table 4.9. The results provide a nice illustration of the fact that some predictions might be more critical than others: The prediction of $r_0^2$ is obviously more equivocal than that of $r_0^1$. In fact, a high degree of plausibility is assigned to $r_0^1 = 0$, whereas $r_0^1 = 1/2$ and $r_0^1 = 1$ are rather unlikely. In the case of $s_0^2$ it is also true that the actual output $1/2$ is the most plausible outcome, yet $r_0^2 = 0$ and $r_0^2 = 1$ are regarded as more or less plausible candidates as well. The different precision of the predictions (see Section 4.6.2 below) is easily understood by inspecting Fig. 4.2: The neighborhood of $s_0^1$ is obviously more homogeneous than that of $s_0^2$. □

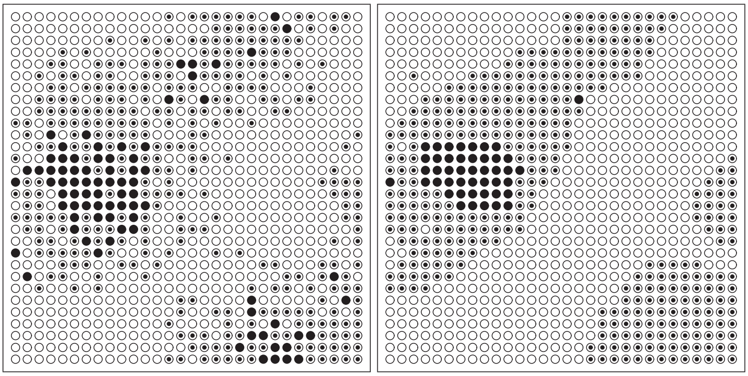


**Fig. 4.11.** Left: The fuzzy concept of Example 4.1 which is also shown in Fig. 4.2. Right: Approximation of this concept (cf. Example 4.23).

Example 4.23. Fig. 4.11 shows a (case-based) approximation $\overline{\varphi}$ of the fuzzy concept of Example 4.1. This approximation, which corresponds to a $\mathcal{S} \longrightarrow \mathcal{R}$ mapping, is again based on a randomly chosen memory $\mathcal{M}$ of size $n = 25$. By considering each instance $(\imath, \jmath)$ as a new input, the approximation (4.34) has been derived for $H = H_{\Sigma}$, where the weights $\alpha_k$ were chosen according to (4.36). Finally, $\overline{\varphi}$ has been determined by

$$\overline{\varphi}(s) = \arg\max_{r \in \mathcal{R}} [\mathsf{Pl}(H, \mathcal{M}, s)](\{r\})$$

for all $s \in \mathcal{S}$, where $\mathsf{Pl}(H, \mathcal{M}, s)$ denotes the plausibility function associated with the belief function $\widehat{\varphi}(s)$ in (4.34). That is, the membership degree of "maximum plausibility" has been chosen as a prediction. As can be seen, this approximation scheme yields a rather good “reconstruction” of the fuzzy concept based on a relatively small number of observations. □

### 4.6.2 More general criteria

In connection with our approach to CBI, a case $\langle s_k, r_k \rangle$ does not provide a simple point estimation, but rather a prediction in the form of a belief function. When rating a case, the following points have thus to be taken into consideration. Firstly, the *correctness* of the prediction does not only depend on the similarity between $r_k$ and the new outcome $r_0$. In fact, the ultimate prediction $\mathsf{Bel}_k = \mathsf{Bel}_k(H, s_0)$ results from the combination of *two* components, namely the outcome $r_k$ and an attached probability measure characterizing the similarity $y = \sigma_{\mathcal{R}}(r_0, r_k)$. That is, $r_k$ is not taken directly as an estimate. Rather, it is interpreted in the context of the probabilistic model. Therefore, $\mathsf{Bel}_k$ might provide a fair prediction even if $r_k$ itself does not: If the probabilistic model is correctly adapted to $\langle s_k, r_k \rangle$, $\mathsf{Bel}_k$ will support outputs which are *not* similar to $r_k$.

Secondly, a prediction in the form of an uncertainty measure suggests not only to rate the correctness of an estimation, but also its *precision.* A prediction specified by the vacuous belief function (i.e., by the mass distribution $\mathsf{m}$ with $\mathsf{m}(\mathcal{R}) = 1$), for instance, might be considered correct, since it makes the true outcome $r_0$ fully plausible. Yet, it does actually not provide any information since it causes the same for all other outcomes as well. This problem (of rating the precision of expert judgements) does also arise in connection with the assessment of human experts which generally specify their incomplete knowledge in the form of, say, probability or possibility distributions [72, 329].

The above considerations suggest to define the weights $\alpha_k$ in accordance with two criteria, namely the correctness and the precision of predictions. As will be seen, both criteria are in fact closely related: In general, a more precise prediction is – in a specific sense – also more correct. It seems, therefore, reasonable to determine the weight of a case in proportion to the precision of the prediction derived from that case.

Assessing the precision of the information obtained from a case $\langle s_k, r_k \rangle$ can be accomplished by quantifying the “amount of uncertainty” related to the prediction $\mathsf{Bel}_k(H, s_0)$ in (4.30). To this end, one can make use of (generalized) measures of uncertainty which have been proposed for belief functions [114, 308, 347, 401]. For a mass distribution $\mathsf{m}$, the following measure defines a reasonable generalization of the *non-specificity* of a set [229, 230]:

$$\mathrm{I}(\mathsf{m}) = \sum_{A \in \mathcal{A}} \mathsf{m}(A) \log_2(|A|),$$

where $\mathcal{A}$ denotes the set of focal elements of $\mathsf{m}$. Thus, a counterpart to (4.36) which is based on the precision of individual estimations can be obtained by letting

$$\alpha_k = \frac{\mathrm{Inf}(\mathsf{m}_k(H, s_0))}{\sum_{\imath=1}^{n} \mathrm{Inf}(\mathsf{m}_\imath(H, s_0))}, \tag{4.37}$$

where $\mathsf{m}_k(H, s_0)$ denotes the mass distribution associated with $\mathsf{Bel}_k(H, s_0)$ and Inf is a specificity measure which is normalized and inversely related to I. When making use of a local $\mathcal{M}$-hypothesis, (4.37) becomes

$$\alpha_k = \frac{\mathrm{Inf}(\mathsf{m}_k(H^{s_k}, s_0))}{\sum_{\imath=1}^{n} \mathrm{Inf}(\mathsf{m}_\imath(H^{s_\imath}, s_0))}.$$

Of course, the precision $\mathrm{I}(m)$ of mass distributions over $\mathcal{R}$ is in general strongly correlated with the precision of the underlying probability measures $H(x)$ from which the mass distributions are derived. The precision of a probability measure $\mu = H(x)$ can be defined, e.g., in terms of the Shannon entropy [338]:

$$\mathrm{I}(\mu) = -\sum_{y \in D_{\mathcal{R}}} \mu(y) \log_2(\mu(y)). \tag{4.38}$$

One might therefore think of replacing (4.37) by

$$\alpha_k = \frac{\mathrm{Inf}(H(\sigma_{\mathcal{S}}(s_0, s_k)))}{\sum_{\imath=1}^{n} \mathrm{Inf}(H(\sigma_{\mathcal{S}}(s_0, s_\imath)))}, \tag{4.39}$$

where Inf is again normalized and inversely related to I as defined in (4.38).

Interestingly enough, (4.38) can also be seen as a measure of the *correctness* of a prediction. In fact, it is readily shown that $\mathrm{I}(H(\sigma_{\mathcal{S}}(s_0, s_k)))$ corresponds to the *expected* correctness of the prediction $\mathsf{Pl}_k(H, s_0)$ if the (in)correctness of this prediction is defined as $\log_2([\mathsf{Pl}_k(H, s_0)](\{\varphi(s_0)\}))$, i.e., as the logarithm of the degree of plausibility assigned to the true outcome $\varphi(s_0)$.

As can be seen, the weighting of a case according to our two criteria, the precision and the (expected) correctness of its prediction, will point in the same direction. Quite often, precision is not only correlated (positively) with correctness, but also with similarity. This observation, which provides a justification for the heuristic approach (4.36), is easily understood by considering two extreme cases: If the similarity $x = \sigma_{\mathcal{S}}(s, s')$ is close to 1, one does expect a large probability for large values $y = \sigma_{\mathcal{R}}(\varphi(s), \varphi(s'))$ and a small probability for small values $y$. The corresponding probability distribution $H(x)$ will hence have a small entropy. This is to be contrasted with a similarity $x$ close to 0, which will hardly allow for making accurate predictions about related similarity degrees $y$ and outcomes $r$. The associated probability distribution might have a stronger tendency toward a uniform distribution[27] and, therefore, will have a larger entropy. Even though rather plausible, the situation will not always be like this, of course. In fact, it is not difficult to construct a counterexample. Anyway, in our example (cf. Table 4.1) the entropy of the distribution $H(x)$ is indeed a decreasing function of $x$ which is very accurately approximated by the mapping $x \mapsto 1.39 - 0.38\,x$ (when leaving the special case $x = 1$ out of account).

[27] Depending on the similarity measure, it will generally reflect but the relative frequency of outcomes.

### 4.6.3 Assessment of individual cases

A weight $a_k$ in (4.36) or (4.39) does not depend on the case $\langle s_k, r_k \rangle$ itself, but only on the similarity between $s_0$ and $s_k$. Even though Section 4.6.2 has given a (heuristic) justification of the case-based determination of weights, these criteria do actually not take the "prediction performance" of an individual case into account. In fact, an observation $\langle s_k, r_k \rangle$ might be rather misleading in the sense that it provides poor predictions, even if $s_k$ is very similar to $s_0$. A simple (but rather drastic) step in this connection is to classify observations into acceptable and non-acceptable ones and to leave the latter out of account. This idea leads to the elimination of what is called *outliers* in statistics and *noisy instances* in instance-based learning [10].

As already mentioned above, the correctness of a prediction $\mathsf{Bel}_k$ does not necessarily require that $r_k$ is close to $r_0$. Rather, it assumes that the similarity $y = \sigma_{\mathcal{R}}(r_0, r_k)$ is accurately specified. Now, recall that the probability measure correctly adapted to the case $\langle s_k, r_k \rangle$ is given by $H_{\Sigma}^{s_k}(\sigma_{\mathcal{S}}(s_0, s_k))$, where $H_{\Sigma}^{s_k}$ is the local measure associated with $s_k$ (cf. Definition 4.12). Consequently, the prediction $\mathsf{Bel}_k$ might be misleading if CBI proceeds from the global PSP $H_{\Sigma}$ resp. a related hypothesis $H$, and the measure $H_{\Sigma}(\sigma_{\mathcal{S}}(s_0, s_k))$ deviates considerably from the local PSP $H_{\Sigma}^{s_k}(\sigma_{\mathcal{S}}(s_0, s_k))$. In fact, relation (4.8) reveals that the $s_k$-PSP can be more or less similar to $H_{\Sigma}$ which represents the "average case" (and is intended to maximize average performance). The more "typical" the input $s_k$ is in this sense, the better the predictions derived from the case $\langle s_k, r_k \rangle$ will be.

Example 4.24. The following table shows the probability distributions $H_{\Sigma}(x)$ and $H_{\Sigma}^{s}(x)$ for the setup $\Sigma = \Sigma_1$, $x = 6/7$, and $s = (2, 8)$ (cf. Example 4.1 and Table 4.1):

| $y$ | 0 | 1/2 | 1 |
|---|---|---|---|
| $[H_{\Sigma_1}(6/7)](y)$ | 0.022 | 0.338 | 0.640 |
| $[H_{\Sigma_1}^{s}(6/7)](y)$ | 0.875 | 0.125 | 0 |

Thus, given a new input $s_0$ directly neighbored to the rather "untypical" input $s$, the measure $H_{\Sigma_1}(6/7)$ suggests $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)) = 1$ and, hence, $\varphi(s_0) = \varphi(s) = 1$ with a probability of 0.64. However, this probability is actually 0, as indicated by $H_{\Sigma_1}^{s}(6/7)$. In fact, it is highly probable that $\sigma_{\mathcal{R}}(\varphi(s), \varphi(s_0)) = 0$, which means that $\varphi(s_0) = 0$. □

If CBI proceeds from a (global) PSP, the above considerations suggest refining the specification of weights $a_k$ by estimating the performance of the cases $\langle s_k, r_k \rangle$. The performance measures can be used, e.g., for adapting the weights which have been determined as a function of similarity. Another possibility is to leave the (case-based) weights as they are, and to *discount* the information provided by a case in accordance with its performance. This can be accomplished by changing a mass distribution $\mathsf{m}_k$ into

$$\mathsf{m}'_k : A \longrightarrow \begin{cases} (1-\lambda_k)\,\mathsf{m}_k(A) & \text{if} \quad A \neq \mathcal{R} \\ (1-\lambda_k)\,\mathsf{m}_k(A) + \lambda_k & \text{if} \quad A = \mathcal{R} \end{cases},$$

where the *discounting factor* $\lambda_k$ is a decreasing function of the performance of the case $\langle s_k, r_k \rangle$.

There are, of course, different approaches to eliciting the performance of a case. Here, let us only indicate one possibility, namely that of learning the prediction performance in a sequence of prediction problems by deriving an estimation of the expected correctness

$$\sum_{s \in \mathcal{S}} \mu_{\mathcal{S}}(s) \, \log_2 \left( [\mathsf{Pl}_k(H,s)](\{\varphi(s)\}) \right) \tag{4.40}$$

associated with a case $\langle s_k, r_k \rangle$.[28] To this end, (4.40) can be approximated by

$$\frac{1}{|\mathcal{T}|} \sum_{\imath=1}^{N} \log_2 \left( [\mathsf{Pl}_k(H,s)](\{\varphi(s_\imath)\}) \right),$$

where $\mathcal{T} = (s_1, \ldots, s_N)$ is a sequence of observed inputs. This approach is in line with the idea of assessing (human) experts by evaluating their performance in the elicitation of a set of "seed" variables [72, 329].

The above discussion reveals that not all cases support the prediction task to the same extent. In fact, the probabilistic model of the CBI hypothesis shows how a case may contradict the similarity-guided extrapolation principle underlying CBI. This way, it can provide the basis of a more sophisticated assessment of cases which goes beyond a simple classification into, say, acceptable and noisy instances. The characterization of cases by means of their local profiles, for example, might be taken as a point of departure for improving heuristic replacement strategies [5, 195, 355, 363] and for complementing other criteria for maintaining optimal memories of cases [253, 282, 356, 357].

Let us make a final remark on a reasonable refinement of the approach discussed so far: Here, we have only been concerned with rating individual cases. In Section 4.5.3 is has however been argued that different cases should not be seen as independent or distinct information sources. Therefore, a case-based inference scheme should take corresponding interdependencies into consideration. As a first step in this direction, we have developed an inference principle that combines (point) predictions from potentially interacting cases by means of the (discrete) Choquet-integral [198]. This method can be seen as a generalization of weighted nearest neighbor estimation. It is not immediately clear, however, how this approach can be further generalized to the setting of this section, where the individual pieces of evidence are belief functions instead of simple point estimations.

[28] This estimation can be seen as a generalization of the *classification record* in instance-based learning, i.e., the number of correct and incorrect classification attempts of each saved instance. Note that the expected correctness depends on the memory $\mathcal{M}$, i.e., it actually assumes $\mathcal{M}$ to be fixed.

An alternative approach could proceed from the specification of dependencies between information sources by means of correlation coefficients, as in random fields [203] or Bayesian methods (with expert opinions modeled as probability distributions [273]). In an extended Bayesian approach, corresponding correlations are taken into consideration when updating a prior belief concerning the new outcome in light of the information provided by different cases.[29] Again, it might be interesting to extend this method to the more general framework of belief functions.

## 4.7 Complex similarity hypotheses

The discussion in Section 4.5 has revealed a main problem of probabilistic inference based on a stochastic model in the form of a PSP, namely the combination of evidence which has been derived from individual cases. This problem is caused by the incomplete knowledge of the dependency structure of these information sources. As shown by Example 4.19, such dependencies cannot be reconstructed without taking further information into account, i.e., information which is not provided by a PSP. The combination of evidence proposed in Section 4.5.3 offers a reasonable solution. Nevertheless, the different possibilities to determine the weights $\alpha_k$ used for combining information sources might be seen with suspicion since they rely on more or less heuristic principles.

There are at least two possibilities for avoiding heuristic reasoning in connection with the transformation of evidence from the similarity level to the instance level. Firstly, one can try to utilize the information available in order to derive approximate results in the form of valid bounds. A corresponding approach to *approximate probabilistic inference* will be discussed in Section 4.8. The second way out is avoiding the combination problem completely. This will be our main concern in the remaining part of this section. Again, there are two possibilities for proceeding. The first approach is to derive inference results from an individual case. This idea has already been discussed in Section 4.5.2 below. The second possibility is the consideration of additional information which is specified by similarity profiles of a more complex nature.

Table 4.10 provides a brief summary of the ideas underlying the probabilistic similarity profiles which have been introduced in Section 4.1. All these profiles realize a $D_{\mathcal{S}} \longrightarrow \mathcal{P}(D_{\mathcal{R}})$ mapping. That is, given the similarity $X$ of two inputs, the respective PSP provides a probabilistic specification of the similarity $Y$ between the associated outcomes. In this section, we shall consider probabilistic profiles which correspond to more general mappings.

[29] The approach pursued in this section is a *direct aggregation* closely related to the consensus method [383]. The direct aggregation of distributions and the Bayesian method are the two basic approaches to combining probability distributions.

| | |
|---|---|
| PSP | Specification of the similarity relation between $S$ and $S_0$, both chosen at random from $\mathcal{S}$. |
| $(n,k)$-PSP | Memory of size $n$, $S$ is taken from the $k$ most similar inputs. |
| $\mathcal{M}$-PSP | Fixed memory $\mathcal{M}$, $S$ is taken from $\mathcal{M}$. |
| $(\mathcal{M},k)$-PSP | Fixed memory $\mathcal{M}$, $S$ is taken from the $k$ most similar inputs in $\mathcal{M}$. |
| $s$-PSP | Profile for a fixed input $s$. |

**Table 4.10.** Basic ideas underlying the probabilistic similarity profiles introduced in Section 4.1.

### 4.7.1 Inference schemes of higher order

The idea of a CBI hypothesis of higher order, i.e., the extension of local inference rules to similarity structures induced by more than two cases, has already been touched upon in Remark 4.20: Given the similarity structure of an extended memory $(\mathcal{M}, s_0)$, the task is to specify a joint probability distribution of the set of similarity degrees

$$\{\sigma_{\mathcal{R}}(\varphi(s), \varphi(S_0)) \,|\, s \in \mathcal{M}^{\downarrow}\}.$$

Admittedly, the number of similarity structures which might occur in connection with a (partial) memory of size $k$ will generally be huge, and this observation remains valid even though this class can be reduced due to symmetry relations and many structures will never occur. It seems, therefore, hopeless to specify a related (probabilistic) hypothesis "by hand." One might still think, however, of employing machine learning methods. This motivates the consideration of the similarity structure introduced in the following definition.

**Definition 4.25 ($n$-PSP of order $k$).** Denote by $\mathcal{Z}_k$ the class of similarity structures $z_S = \mathsf{SST}(\mathcal{M}^{s_0}, s_0)$, where $s_0 \in \mathcal{S}$, and $\mathcal{M}^{s_0}$ denotes the $s_0$-*ordered* version of a memory $\mathcal{M}$. That is, $\mathcal{M}^{s_0}$ is a permutation of $\mathcal{M}$ such that the vector

$$(x_{01}, \ldots, x_{0k}, x_{12}, x_{13}, \ldots, x_{k-1,k}, y_{12}, y_{13}, \ldots, y_{k-1,k}),$$

with $x_{\imath\jmath} = \sigma_{\mathcal{S}}(s_\imath, s_\jmath)$ and $y_{\imath\jmath} = \sigma_{\mathcal{R}}(\varphi(s_\imath), \varphi(s_\jmath))$ for $0 \leq \imath < \jmath \leq k$, is minimal with respect to a lexicographic order. Now, consider a CBI setup and let $\mathcal{M} \sim (\mu_{\mathcal{S}\times\mathcal{R}})^n$, $S_0 \sim \mu_{\mathcal{S}}$,

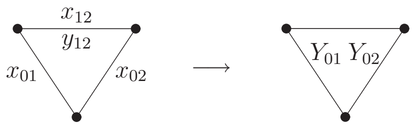
$$\mathcal{M}_k = \mathcal{N}_k(\mathcal{M}, S_0) = \big(\langle S_1, \varphi(S_1)\rangle, \ldots, \langle S_k, \varphi(S_k)\rangle\big)\,.$$

Moreover, let $Z_S = \mathsf{SST}(\mathcal{M}_k^{S_0}, S_0)$, $Y_{0\jmath} = \sigma_{\mathcal{R}}(\varphi(S_0), \varphi(S_\jmath))$, and $Y = (Y_{01}, \ldots, Y_{0k})$. The $n$-PSP of order $k$ is the mapping

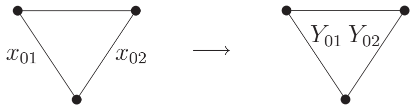
$$H_{\Sigma}^{(n,k)} : \mathcal{Z}_k \longrightarrow \mathcal{P}\big((D_{\mathcal{R}})^k\big)\,,\; z_S \mapsto \mu_{Y\,|\,(Z_S = z_S)}.$$

That is, $[H_{\Sigma}^{(n,k)}(z_S)](y_1, \ldots, y_k)$ is the probability that $\sigma_{\mathcal{R}}(\varphi(S_0), \varphi(S_\jmath)) = y_\jmath$ for all $1 \leq \jmath \leq k$, given the information provided by the similarity structure $\mathsf{SST}(\mathcal{M}_k^{S_0}, S_0)$. □

The idea of *averaging* over similarity profiles has already been mentioned in Section 4.1. Equations (4.47) and (4.48) reveal that such an averaging corresponds to the weighted aggregation of local inference rules. The same principle can obviously be applied in connection with profiles of higher order. A local inference rule associated with a PSP of order 2, for instance, assigns to each similarity structure $z_S = (x_{01}, x_{02}, x_{12}, y_{12})$ the (conditional) probability distribution of the similarity vector $(Y_{01}, Y_{02})$ and, hence, can be illustrated graphically as follows:



Now, suppose that we ignore the similarity relation between the observed cases, expressed by the variables $x_{12}$ and $y_{12}$. That is, we deduce the probability distribution of the random vector $(Y_{01}, Y_{02})$ from the partial similarity structure $(x_{01}, x_{02})$:



The resulting inference rule corresponds to the weighted aggregation of those rules the precedent of which can be "matched" with the partial structure $z_S = (x_{01}, x_{02}, \cdot, \cdot)$. On the one hand, such an aggregation is obviously connected with a loss of information. On the other hand, the complexity of inference schemes can be reduced considerably this way. In fact, ignoring the similarity relations between the cases in the memory reduces the size of a similarity structure from $k^2$ to $k$ and, hence, the number of different structures to $(D_{\mathcal{S}})^k$. Therefore, the following definition seems reasonable.

**Definition 4.26 (partial $n$-PSP of order $k$).** Denote by $\mathcal{Z}_k$ the class of partial similarity structures $z_S = \mathsf{pSST}(\mathcal{M}^{s_0}, s_0)$, where $s_0 \in \mathcal{S}$ and $\mathcal{M}$ is a memory of size $k$. Consider a CBI setup and let $\mathcal{M} \sim (\mu_{\mathcal{S}\times\mathcal{R}})^n$, $S_0 \sim \mu_{\mathcal{S}}$,

$$\mathcal{M}_k = \mathcal{N}_k(\mathcal{M}, S_0) = \big(\langle S_1, \varphi(S_1)\rangle, \ldots, \langle S_k, \varphi(S_k)\rangle\big)\,.$$

Moreover, $Z_S = \mathsf{pSST}(\mathcal{M}_k^{S_0}, S_0)$, $Y_{0\jmath} = \sigma_{\mathcal{R}}(\varphi(S_0), \varphi(S_\jmath))$, and $Y = (Y_{01}, \ldots, Y_{0k})$. The mapping

$$H_{\Sigma}^{(n,k)} : \mathcal{Z}_k \longrightarrow \mathcal{P}\big((D_{\mathcal{R}})^k\big)\,,\; z_S \mapsto \mu_{Y\,|\,(Z_S = z_S)}.$$

is called the partial $n$-PSP of order $k$. □

The assumption of a fixed memory $\mathcal{M}$ underlying CBI leads to the definition of a (partial) $\mathcal{M}$-PSP of order $k$. The idea of considering only *partial* similarity structures arises quite naturally in this context. In connection with the $\mathcal{M}$-PSP of order $n = \operatorname{card}(\mathcal{M})$, for instance, it is not necessary to take the similarity relations between cases in $\mathcal{M}$ into account since these relations do not change.

**Definition 4.27 ($\mathcal{M}$-PSP of order $k$).** Consider a CBI setup $\Sigma$ with a *fixed* memory $\mathcal{M}$, where $\operatorname{card}(\mathcal{M}) \geq k$. For $S_0 \sim \mu_{\mathcal{S}}$, let

$$\mathcal{M}_k = \mathcal{N}_k(\mathcal{M}, S_0) = \big(\langle s_1, \varphi(s_1)\rangle, \ldots, \langle s_k, \varphi(s_k)\rangle\big) .$$

Moreover, $Z_S = \mathsf{SST}(\mathcal{M}_k^{S_0}, S_0)$, $Y_{0\jmath} = \sigma_{\mathcal{R}}(\varphi(S_0), \varphi(s_\jmath))$, and $Y = (Y_{01}, \ldots, Y_{0k})$. The mapping

$$H_\Sigma^{(n,k)} : \mathcal{Z}_k \longrightarrow \mathcal{P}((D_{\mathcal{R}})^k) \, , \; z_S \mapsto \mu_{Y|(Z_S=z_S)}$$

is called the $\mathcal{M}$-PSP of order $k$. □

**Definition 4.28 (partial $\mathcal{M}$-PSP of order $k$).** Consider a CBI setup $\Sigma$ with a *fixed* memory $\mathcal{M}$, where $\operatorname{card}(\mathcal{M}) \geq k$. For $S_0 \sim \mu_{\mathcal{S}}$, let

$$\mathcal{M}_k = \mathcal{N}_k(\mathcal{M}, S_0) = \big(\langle s_1, \varphi(s_1)\rangle, \ldots, \langle s_k, \varphi(s_k)\rangle\big) .$$

Moreover, $Z_S = \mathsf{pSST}(\mathcal{M}_k^{S_0}, S_0)$, $Y_{0\jmath} = \sigma_{\mathcal{R}}(\varphi(S_0), \varphi(s_\jmath))$, and $Y = (Y_{01}, \ldots, Y_{0k})$. The mapping

$$H_\Sigma^{(n,k)} : \mathcal{Z}_k \longrightarrow \mathcal{P}((D_{\mathcal{R}})^k) \, , \; z_S \mapsto \mu_{Y|(Z_S=z_S)}$$

is called the partial $\mathcal{M}$-PSP of order $k$. □

Consider a CBI problem $\langle \Sigma, s_0 \rangle$, with $\mathcal{M}$ being the memory of $\Sigma$. A probabilistic inference scheme (of higher order) can now be realized in the form

$$\eta = \sigma_{\mathcal{R}}^{(-1)} \big(\mathcal{N}_k(\mathcal{M}, s_0), H(z_S)\big) \, , \tag{4.41}$$

where $H$ is a hypothesis related to a (partial) $n$-PSP or $\mathcal{M}$-PSP of order $k$ and $z_S$ is the (partial) similarity structure defined by $s_0$ and the $k$-selection $\mathcal{N}_k(\mathcal{M}, s_0)$. The transformation $\sigma_{\mathcal{R}}^{(-1)}$ can be realized as in Section 4.5.1. In fact, the $k$ cases $\mathcal{N}_k(\mathcal{M}, s_0)$ can now be considered as one information source. They induce an imperfect specification $\Gamma = (\gamma, p_C)$, where

$$\begin{array}{rcl} C & = & (D_{\mathcal{R}})^k, \\ p_C & = & H(z_S), \\ \gamma(c) & = & \bigcap_{1 \leq k \leq n} \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), c_k) \end{array}$$

for all $c = (c_1, \ldots, c_n) \in C$. This is nothing else than the imperfect specification (4.27) which has already been proposed in Section 4.5.3 as an “ideal” combination of evidence but which could not be realized due to the fact that the measure $\mu$ was not known. This measure is now given by $H(z_S)$.

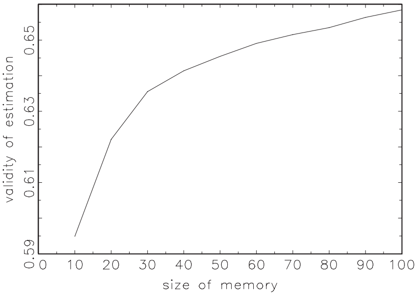


**Fig. 4.12.** Probability of a correct estimation based on the partial $\mathcal{M}$-PSP of order 3, where $\mathcal{M}$ is a memory of size $n$ (cf. Example 4.29).

EXAMPLE 4.29. For the CBI setup $\Sigma_3$ we have carried out an experimental study in which we were interested in the partial $\mathcal{M}$-PSP of order 3 for memories of size $n$. Based on such a profile, the following inference strategy is realized: Given a new input $s_0$, a measure $\eta \in \mathcal{F}(\mathcal{R})$ is derived according to (4.41), and the outcome $\varphi(s_0)$ is estimated by $\widehat{\varphi}(s_0) = \arg\max_{r \in \mathcal{R}} \eta(r)$.[30] For a certain memory $\mathcal{M}$ let $p_{\mathcal{M}} = \mathbb{P}(\widehat{\varphi}(S_0) = \varphi(S_0) \,|\, \mathcal{M})$ denote the probability of a correct estimation. From a large sample of memories we have derived an approximation of the expected value

$$\mathbb{E}(p_{\mathcal{M}}) = \sum_{\mathcal{M} \subset \mathcal{S} \times \mathcal{R}} p_{\mathcal{M}} \cdot \mu_{\mathcal{S} \times \mathcal{R}}(\mathcal{M})$$

of this probability. Fig. 4.12 shows the results for different values of $n$. □

### 4.7.2 Partially admissible profiles

In Section 3.2, we argued that the inference scheme (3.2) provides correct predictions if the similarity hypothesis $h$ is admissible. More precisely, admissibility of $h$ is a *sufficient* condition. It is, however, not a *necessary* one. In fact, the prediction (3.2) is already correct if $h$ is *partially admissible* in the sense that it obtains for the subset of cases $\langle s_k, r_k \rangle$ $(0 \leq k \leq n)$ actually involved in the inference process:

$$\forall\, 1 \leq \imath \leq n \,:\, h(\sigma_{\mathcal{S}}(s_0, s_\imath)) \leq \sigma_{\mathcal{R}}(\varphi(s_0), \varphi(s_\imath)). \tag{4.42}$$

We say that a hypothesis $h : [0,1] \longrightarrow [0,1]$ is admissible with respect to a sequence $(s_0, s_1, \ldots, s_n) \in \mathcal{S}^{n+1}$ if (4.42) holds true.

[30] Ties are broken by coin flipping.

**Definition 4.30 ($\mathcal{H}$-PSP).** Consider a setup $\Sigma$ and let $\mathcal{H}$ be a set of profiles $h : D_{\mathcal{S}} \longrightarrow [0,1]$ such that $(\mathcal{H}, \leq)$ is a complete order. The mapping

$$H_{\Sigma} : \mathfrak{N} \times \mathcal{H} \longrightarrow [0,1] \tag{4.43}$$

with

$$H_{\Sigma}(n,h) \overset{\mathrm{df}}{=} (\mu_{\mathcal{S}})^{n+1} \left( \{ s \in \mathcal{S}^{n+1} \,|\, h \text{ is admissible w.r.t. } s \} \right)$$

is called the $\mathcal{H}$-PSP of $\Sigma$. For each $n \in \mathfrak{N}$ and hypothesis $h \in \mathcal{H}$, the profile $H_{\Sigma}$ specifies the probability $H_{\Sigma}(n,h)$ that $h$ is admissible with respect to a (random) sequence $(S_0, S_1, \ldots, S_n)$ of $n$ inputs. □

**Definition 4.31 ($\mathcal{H}$-probabilistic similarity hypothesis).** Let $\mathcal{H}$ be a set of hypotheses $h : [0,1] \longrightarrow [0,1]$ completely ordered by $\leq$. A mapping $H : \mathfrak{N} \times \mathcal{H} \longrightarrow [0,1]$ such that

$$\forall\, h, h' \in \mathcal{H} \, : h \leq h' \Rightarrow H(n,h) \geq H(n,h'),$$
$$\forall\, n, n' \in \mathfrak{N} \, : n \leq n' \Rightarrow H(n,h) \geq H(n,h')$$

is called an $\mathcal{H}$-probabilistic similarity hypothesis. A hypothesis $H$ is admissible with respect to a setup $\Sigma$ if $H(n,h) \leq H_{\Sigma}(n,h)$ for all $n \in \mathfrak{N}$ and $h \in \mathcal{H}$. $H$ is called stronger than a hypothesis $H'$ if $H' \leq H$ and $H \not\leq H'$. □

For a CBI problem $\langle \Sigma, s_0 \rangle$ and a hypothesis $h \in \mathcal{H}$, the prediction

$$\varphi(s_0) \in \varphi_{h,\mathcal{M}}(s_0) \overset{\mathrm{df}}{=} \bigcap_{1 \leq k \leq n} \mathcal{N}_{h(\sigma_{\mathcal{S}}(s_0, s_k))}(\varphi(s_k))$$

is correct with probability $H_{\Sigma}(\mathrm{card}(\mathcal{M}), h)$. That is, the prediction

$$\varphi_{H_{\Sigma},\mathcal{M}}(s_0) \overset{\mathrm{df}}{=} \{ \varphi_{h,\mathcal{M}}(s_0) \,|\, h \in \mathcal{H} \}$$

defines a class of confidence regions for the unknown outcome $\varphi(s_0)$.

The concept of an $\mathcal{H}$-PSP can be generalized in accordance with the ideas of utilizing a fixed memory $\mathcal{M}$, or of basing CBI on the $k$ most similar cases (cf. Section 3.1). Assuming a fixed memory $\mathcal{M}$ simplifies the specification of a hypothesis since the $\mathcal{H}$-PSP does no longer depend on the parameter $n$. That is, the partial admissibility of a hypothesis depends only on the new input $s_0$, and (4.43) becomes a mapping

$$H_{\Sigma}^{\mathcal{M}} : \mathcal{H} \longrightarrow [0,1].$$

Again, a hypothesis related to an $\mathcal{H}$-PSP may originate from different sources. A reasonable idea is to take a parameterized class $\mathcal{H}$ of hypotheses as a point of departure and to learn the probabilities $H_{\Sigma}(n,h)$ associated with different hypotheses $h \in \mathcal{H}$ from observed data. Based on a sequence of CBI problems, the probability of a hypothesis $h$ to be admissible can simply be estimated from the relative frequency of correct predictions, or a prior estimation thereof can be revised accordingly.

| | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.00 | 0.02 | 0.05 | 0.11 | 0.20 | 0.27 | 0.33 | 0.36 |
| 1/7 | | 0.28 | 0.36 | 0.48 | 0.59 | 0.69 | 0.75 | 0.78 |
| 2/7 | | | 0.39 | 0.52 | 0.65 | 0.75 | 0.82 | 0.86 |
| 3/7 | | | | 0.55 | 0.68 | 0.80 | 0.87 | 0.91 |
| 4/7 | | | | | 0.69 | 0.82 | 0.90 | 0.95 |
| 5/7 | | | | | | 0.84 | 0.92 | 0.98 |
| 6/7 | | | | | | | 0.94 | 0.99 |
| 1 | | | | | | | | 1 |

**Table 4.11.** Probability of the admissibility of hypotheses $h_{u,v}$ (cf. Example 4.32). Rows and columns correspond to the parameters $u$ and $v$, respectively.

EXAMPLE 4.32. For the setup $\Sigma_3$ (cf. Example 4.1) we have specified a class of (strict) hypotheses by means of two parameters $u, v \in \{0, 1/7, \ldots, 1\}$, where $u \leq v$:

$$h_{u,v} : D_{\mathcal{S}} \longrightarrow D_{\mathcal{R}} \, , \; x \mapsto \begin{cases} 0 & \text{if} \quad x < u \\ 1/2 & \text{if} \quad u \leq x \leq v \\ 1 & \text{if} \quad v \leq x \end{cases} . \tag{4.44}$$

Table 4.11 shows the probability of the admissibility of these hypotheses in connection with a memory of size $n = 20$.[31] A subset $\mathcal{H}$ of hypotheses (4.44) completely ordered by $\leq$ can be chosen in order to define (4.43) for $n = 20$. □

## 4.8 Approximate probabilistic inference

### 4.8.1 Generalized uncertainty measures and profiles

A value $h_\Sigma(x)$ of a similarity profile can be written as

$$\begin{aligned} h_\Sigma(x) &= \inf_{r \in \mathcal{R}} h'_\Sigma(x, r) & (4.45) \\ &= \inf_{r \in \mathcal{R}, s \in \mathcal{S}} h''_\Sigma(x, r, s), & (4.46) \end{aligned}$$

where the (more specific) profile $h'_\Sigma$ in (4.45) is defined on $D_{\mathcal{S}} \times \mathcal{R}$ by

$$h'_\Sigma : (x, r) \mapsto \inf_{s, s' \in \mathcal{S} : \sigma_{\mathcal{S}}(s, s') = x, \varphi(s) = r} \sigma_{\mathcal{R}}(\varphi(s), \varphi(s')),$$

and $h''_\Sigma : D_{\mathcal{S}} \times \mathcal{R} \times \mathcal{S} \longrightarrow D_{\mathcal{R}}$ is defined correspondingly. The profile $h'_\Sigma$ can be associated with rules of the following form: "If the similarity between two inputs is $x$ and the outcome of the first input is $r$, then the similarity between the associated outputs is at least $h'_\Sigma(x, r)$." Equations (4.45) and (4.46) show that a similarity profile provides a lower bound to the respective underlying class of

[31] Actually, these probabilities are estimations from a large number of randomly generated memories.

more specific profiles. Within the probabilistic setting of this chapter, "bounding" is replaced by "averaging." Consider the probabilistic counterparts to (4.45) and (4.46) as an example, i.e., profiles of the form $H''_\Sigma : D_\mathcal{S} \times \mathcal{R} \times \mathcal{S} \longrightarrow \mathcal{P}(D_\mathcal{R})$ and $H'_\Sigma : D_\mathcal{S} \times \mathcal{R} \longrightarrow \mathcal{P}(D_\mathcal{R})$. The distribution $H'_\Sigma(x,r)$, for instance, specifies the similarity $\sigma_\mathcal{R}(\varphi(S), \varphi(S'))$ given that $\sigma_\mathcal{S}(S,S') = x$ and $\varphi(S) = r$ for two inputs $S, S' \in \mathcal{S}$. We have

$$\begin{aligned} H'_\Sigma(x,r) &\propto \sum_{s \in \mathcal{S}} \alpha(x,r,s) \cdot H''_\Sigma(x,r,s), & (4.47) \\ H_\Sigma(x) &\propto \sum_{r \in \mathcal{R}} \beta(x,r) \cdot H'_\Sigma(x,r), & (4.48) \end{aligned}$$

for $x \in D_\mathcal{S}, r \in \mathcal{R}$ and $x \in D_\mathcal{S}$, respectively, where

$$\begin{aligned} \beta(x,r) &= \mu_\mathcal{R}(r) \cdot \mu_{X|(R=r)}(x), \\ \alpha(x,r,s) &= \mu_\mathcal{S}(s) \cdot \mu_{R|(S=s)}(r) \cdot \mu_{X|(R=r,S=s)}(x). \end{aligned}$$

In other words, the PSP $H_\Sigma$ is a weighted average of the more specific profiles $H'_\Sigma$ and $H''_\Sigma$, respectively.

The derivation of bounds instead of averages of similarity profiles may become interesting within the probabilistic setting as well. We might utilize, for instance, upper probabilities in the inference rules instead of probability measures. For example, rather than deriving $\mu_{Y|(X=x)}$ according to (4.47) and (4.48), one could aggregate the information contained in the measures on the right-hand side by means of upper envelopes [83]

$$\eta_{Y|(X=x)} \overset{\mathrm{df}}{=} \bigvee_{r \in \mathcal{R}, s \in \mathcal{S}} \mu_{Y|(X=x,R=r,S=s)}. \tag{4.49}$$

Note that (4.49) is a natural generalization of the constraint-based approach of Section 3, where

$$[h_\Sigma(x), 1] = \bigcup_{r \in \mathcal{R}, s \in \mathcal{S}} [h''_\Sigma(x,r,s), 1] \tag{4.50}$$

for all $x \in D_\mathcal{S}$.

In connection with the derivation of upper bounds, the concept of a *normalized uncertainty measure* (cf. Section 4.5) turns out to be useful. Denote by $\mathcal{F}(\Omega, \mathcal{A})$ (or simply $\mathcal{F}(\Omega) = \mathcal{F}(\Omega, 2^\Omega)$) the class of normalized uncertainty measures defined on a measurable space $(\Omega, \mathcal{A})$, i.e., the class of measures $\eta : \mathcal{A} \longrightarrow [0,1]$ satisfying

- $\eta(\emptyset) = 0$, $\eta(\Omega) = 1$,
- $\forall\, A, B \in \mathcal{A} : A \subset B \Rightarrow \eta(A) \leq \eta(B)$.

**Definition 4.33 (generalized similarity profile).** It was already mentioned that the PSP $H_\Sigma$ is a weighted average of the more specific profiles (4.47) resp. (4.48). A mapping $G_\Sigma : D_\mathcal{S} \longrightarrow \mathcal{F}(D_\mathcal{R})$ which has been derived by means of an alternative aggregation

$$G_\Sigma(x) = A\left(\{\mu_{Y|(X=x,R=r,S=s)} \,|\, s \in \mathcal{S}, r \in \mathcal{R}\}\right)$$

is called a generalized similarity profile (GSP) of the setup $\Sigma$. □

**Definition 4.34 (generalized similarity hypothesis).** A generalized similarity hypothesis is identified by a mapping $G : D_\mathcal{S} \longrightarrow \mathcal{F}(D_\mathcal{R})$. Let $\Sigma$ be a CBI setup with PSP $H_\Sigma$. The hypothesis $G$ is admissible if $H_\Sigma(x)$ dominates $G(x)$ for all $x \in D_\mathcal{S}$ in the sense that $F_{H_\Sigma(x)} \leq F_{G(x)}$. (For all $\eta \in \mathcal{F}(D_\mathcal{R})$, $F_\eta : D_\mathcal{R} \longrightarrow [0,1]$ denotes the mapping $x \mapsto \eta(D_\mathcal{R} \cap [0,x])$.) $G$ is called a *strict* generalized hypothesis if

$$\forall\, x, x' \in D_\mathcal{S} \,:\, x' < x \Rightarrow F_{G(x)} \leq F_{G(x')}.$$

A hypothesis $G'$ satisfying $F_{G'(x)} \leq F_{G(x)}$ for all $x \in D_\mathcal{S}$ and $F_{G(x_0)} \not\leq F_{G'(x_0)}$ for at least one $x_0 \in D_\mathcal{S}$ is called *stronger* than $G$. □

Observe that we compare the cumulative distribution function of a probability measure to that of a generalized measure in Definition 4.34. It should be mentioned, therefore, that a GSP has been introduced with a probabilistic interpretation of a generalized measure in mind. Particularly, a measure $\eta \in \mathcal{F}(\Omega, \mathcal{A})$ might be thought of as an *upper probability* in the sense that it provides an estimation of a probability $\mu \in \mathcal{P}(\Omega, \mathcal{A})$ in the form of upper bounds: $\eta(A) \geq \mu(A)$ for all $A \in \mathcal{A}$. Likewise, $\eta$ might define an *upper envelope* of a family $\mathcal{C}$ of (probability) measures, i.e., $\eta(A) = \sup_{\mu \in \mathcal{C}} \mu(A)$ for all $A \in \mathcal{A}$. Thus, a generalized hypothesis $G$ is associated with implications of the form

$$\sigma_\mathcal{S}(S, S') = x \Rightarrow \mu_{Y|(X=x)} \leq G(x),$$

whereas a corresponding probabilistic hypothesis $H$ defines implications

$$\sigma_\mathcal{S}(S, S') = x \Rightarrow \mu_{Y|(X=x)} = H(x).$$

Seen from this perspective, a generalized hypothesis $H$ is again admissible if it is "pessimistic" enough: For all $y \in [0,1]$, the probability that the degree of similarity between two outcomes is equal to or less than $y$ is not under-estimated by $H$.

Example 4.35. Consider again the CBI setup $\Sigma_3$ which has been introduced in Example 4.1. Table 4.12 shows the probabilistic profiles $H'_{\Sigma_3}(x,r) = \mu_{Y|(X=x,R=r)}$ for $x \in D_{\mathcal{S}_3}$ and $r \in \mathcal{R} = \{0, 1/2, 1\}$. Table 4.13 shows the values of the upper envelope $\eta(x) = \bigvee_{r \in \mathcal{R}} \mu_{Y|(X=x,R=r)}$. This measure defines a generalized similarity profile. □

| $\mu_{Y \mid (X=x,R=r)}(0)$ | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.53 | 0.59 | 0.62 | 0.66 | 0.69 | 0.73 | 0.75 | 1.00 |
| 1/2 | 0.35 | 0.37 | 0.39 | 0.42 | 0.47 | 0.52 | 0.54 | 1.00 |
| 1 | 0.06 | 0.13 | 0.18 | 0.24 | 0.33 | 0.41 | 0.47 | 1.00 |

| $\mu_{Y \mid (X=x,R=r)}(1/2)$ | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.37 | 0.34 | 0.32 | 0.29 | 0.27 | 0.25 | 0.23 | 0.00 |
| 1/2 | 0.65 | 0.63 | 0.61 | 0.57 | 0.53 | 0.48 | 0.46 | 0.00 |
| 1 | 0.34 | 0.47 | 0.49 | 0.50 | 0.46 | 0.43 | 0.41 | 0.00 |

| $\mu_{Y \mid (X=x,R=r)}(1)$ | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.10 | 0.07 | 0.06 | 0.05 | 0.04 | 0.03 | 0.02 | 0.00 |
| 1/2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1 | 0.60 | 0.41 | 0.33 | 0.26 | 0.21 | 0.15 | 0.13 | 0.00 |

**Table 4.12.** Probabilistic profiles $H'_{\Sigma_3}$ (cf. Example 4.35). Columns correspond to values of $x$, rows to values of $r$.

| $x$ | 0 | 1/7 | 2/7 | 3/7 | 4/7 | 5/7 | 6/7 | 1 |
|---|---|---|---|---|---|---|---|---|
| $\eta_{Y \mid (X=x)}(0)$ | 0.60 | 0.41 | 0.33 | 0.26 | 0.21 | 0.15 | 0.13 | 0.00 |
| $\eta_{Y \mid (X=x)}(1/2)$ | 0.65 | 0.63 | 0.61 | 0.57 | 0.53 | 0.48 | 0.46 | 0.00 |
| $\eta_{Y \mid (X=x)}(1)$ | 0.53 | 0.59 | 0.62 | 0.66 | 0.69 | 0.73 | 0.75 | 1.00 |

**Table 4.13.** Generalized similarity profile of the setup $\Sigma_3$ (cf. Example 4.35).

### 4.8.2 An approximate inference scheme

We shall now take up the idea of the likelihood approach (4.4) for estimating the unknown outcome $r_0 = \varphi(s_0)$. However, due to the problem of combining probabilistic evidence (cf. Section 4.5) we will derive not the likelihood itself but only an upper approximation thereof.

Consider a CBI problem $\langle \Sigma, s_0 \rangle$. Proceeding from the outcome structure $z_O = \mathsf{OST}(\mathcal{M}, s_0)$ we can write the likelihood function (4.4) as follows:

$$\lambda(r) \overset{\text{df}}{=} \mathbb{P}\left(Z_O = z_O \,|\, R_0 = r\right) \tag{4.51}$$

for all $r \in \mathcal{R}$. Now, for $1 \leq k \leq n$ we have

$$\begin{aligned} \lambda(r) &= \mathbb{P}\left(X_{0k} = x_{0k}, R_k = r_k \,|\, R_0 = r\right) \cdot \lambda'_k(r) \\ &\leq \mathbb{P}\left(X_{0k} = x_{0k}, R_k = r_k \,|\, R_0 = r\right), \end{aligned}$$

where

$$\lambda'_k(r) \overset{\text{df}}{=} \mathbb{P}\left(Z'_k = z'_k \,|\, R_0 = r, X_{0k} = x_{0k}, R_k = r_k\right)$$

with $Z'_k = Z_O - \{X_{0k}, R_k\}$ and $z'_k = z_O - \{x_{0k}, r_k\}$.

Since $(R_0 = r \wedge R_k = r_k) \Rightarrow Y_{0k} = \sigma_{\mathcal{R}}(r, r_k)$, we also have

$$\begin{aligned} &\mathbb{P}\left(X_{0k} = x_{0k}, R_k = r_k \,|\, R_0 = r\right) = \\ &\quad = \mathbb{P}\left(X_{0k} = x_{0k} \,|\, R_0 = r\right) \cdot \mathbb{P}\left(R_k = r_k \,|\, R_0 = r,\ X_{0k} = x_{0k}\right) \\ &\quad \leq \mathbb{P}\left(Y_k = \sigma_{\mathcal{R}}(r, r_k) \,|\, R_0 = r,\ X_{0k} = x_{0k}\right). \end{aligned}$$

Thus, we finally derive the estimation

$$\begin{aligned} \lambda(r) &\leq \min_{1\leq k\leq n} \mathbb{P}\left(Y_k = \sigma_{\mathcal{R}}(r,r_k) \,|\, R_0 = r,\, X_{0k} = x_{0k}\right) \\ &= \min_{1\leq k\leq n} \mu_{Y|(X=x_{0k},R=r)}(\sigma_{\mathcal{R}}(r,r_k)). \end{aligned} \tag{4.52}$$

For $x \in D_{\mathcal{S}}$, let $\eta_{Y|(X=x)}$ denote an upper approximation of the class

$$\{\mu_{Y|(X=x,R=r)} \,|\, r \in \mathcal{R}\} \tag{4.53}$$

of measures. That is,

$$\forall\, x \in D_{\mathcal{S}} \;:\; \eta_{Y|(X=x)} \geq \bigvee_{r\in\mathcal{R}} \mu_{Y|(X=x,R=r)}.$$

These measures, which define a generalized similarity profile $G : x \mapsto \eta_{Y|(X=x)}$, are upper approximations of the probability measures $H_\Sigma(x)$ $(x \in D_{\mathcal{S}})$. In other words, $G$ is an upper approximation of $H_\Sigma$. We now obtain the following approximation of (4.51):

$$\lambda(r) \leq \min_{1\leq k\leq n} \eta_{Y|(X=\sigma_{\mathcal{S}}(s_0,s_k))}(\sigma_{\mathcal{R}}(r,r_k)).$$

**Proposition 4.36.** Consider a CBI problem $\langle\Sigma, s_0\rangle$ and suppose the generalized hypothesis $G$ to satisfy $G(x) \geq \bigvee_{r\in\mathcal{R}} \mu_{Y|(X=x,R=r)}$ for all $x \in D_{\mathcal{S}}$. The function

$$\overline{\lambda} : r \mapsto \min_{\langle s,\varphi(s)\rangle\in\mathcal{M}} [G(\sigma_{\mathcal{S}}(s,s_0))](\sigma_{\mathcal{R}}(\varphi(s),r)) \tag{4.54}$$

is an upper approximation of the likelihood function (4.51), i.e., $\lambda(r) \leq \overline{\lambda}(r)$ for all $r \in \mathcal{R}$. □

The result of Proposition 4.36 can directly be used for putting the idea of *approximate* case-based inference (cf. page 146) into action. Given the CBI problem $\langle\Sigma, s_0\rangle$ and a related generalized similarity hypothesis, i.e., a mapping $G$ such that $G(x)$ is a normalized uncertainty measure for all $x \in S_{\mathcal{S}}$, approximate probabilistic CBI comes down to predicting $\varphi(s_0)$ according to (4.54).

Example 4.37. Reconsider Example 4.35 where we have derived a generalized similarity profile related to the CBI setup $\Sigma_3$. This profile, which is shown in Table 4.13, satisfies the assumptions of Proposition 4.36. In Example 4.13, we have considered the CBI problem $\langle\Sigma_3, s_0\rangle$ with $\mathcal{M} = (\langle(3,14),1/2\rangle, \langle(4,17),1/2\rangle)$ and $s_0 = (5,17)$. Applying the approximate reasoning scheme (4.54) to this problem, we obtain the following results:

| $r$ | 0 | 1/2 | 1 |
|---|---|---|---|
| $\overline{\lambda}(r)$ | 0.46 | 0.75 | 0.46 |

Since this rating of outcomes is expressed in terms of (approximations of) degrees of likelihood it is not directly comparable to the probabilities $\mathbb{P}(R_0 = r | Z_O = z_O)$ tabulated in Example 4.13. Nevertheless, the results are qualitatively similar and $r = 1/2$ is promoted as the most likely outcome in both cases. □

An alternative approach to approximating (4.4) is to proceed from the *case structure* $z_C = \mathsf{CST}(\mathcal{M}, s_0)$ instead of the outcome structure of a CBI problem $\langle \Sigma, s_0 \rangle$.

**Definition 4.38 (case structure).** Let $\Sigma$ be a CBI setup with $\mathcal{M}$ being the associated memory (2.29) of cases and let $s_0$ be a new input. The set of values

$$\mathsf{CST}(\mathcal{M}, s_0) \overset{\text{df}}{=} \mathsf{OST}(\mathcal{M}, s_0) \cup \{s_\jmath \,|\, 1 \leq \jmath \leq n\}$$

(together with $(h_\Sigma, \sigma_\mathcal{S}, \sigma_\mathcal{R})$) defines the case structure of the CBI problem $\langle \Sigma, s_0 \rangle$. □

Making use of the information provided by a case structure, (4.51) becomes

$$\lambda(r) \overset{\text{df}}{=} \mathbb{P}\left(Z_C = z_C \,|\, S_0 = s_0, R_0 = r\right) \tag{4.55}$$

for all $r \in \mathcal{R}$. We have

$$\begin{aligned} \lambda(r) &= \mathbb{P}\left(Z'_C = z'_C \,|\, S_0 = s_0, R_0 = r\right) \cdot \lambda'(r) \\ &\leq \mathbb{P}\left(Z'_C = z'_C \,|\, S_0 = s_0, R_0 = r\right), \end{aligned} \tag{4.56}$$

where

$$\lambda'(r) = \mathbb{P}\left(Z''_C = z''_C \,|\, S_0 = s_0, R_0 = r, Z'_C = z'_C\right)$$

with $z'_C = z_C - z''_C$ and $z''_C = z_C - \{s_k, r_k, x_{0k} \,|\, 1 \leq k \leq n\}$. Since the random variables $Z_k = (S_k, R_k, X_{0k})$ $(1 \leq k \leq n)$ are *conditionally independent* given $(S_0, R_0)$, (4.56) becomes

$$\begin{aligned} \lambda(r) &= \lambda'(r) \cdot \prod_{k=1}^{n} \mathbb{P}\left(S_k = s_k, R_k = r_k, X_{0k} = x_{0k} \,|\, S_0 = s, R_0 = r\right) \\ &\leq \prod_{k=1}^{n} \mathbb{P}\left(R_k = r_k, X_{0k} = x_{0k} \,|\, S_0 = s_0, R_0 = r\right). \end{aligned}$$

Again, using $(R_0 = r \wedge R_k = r_k) \Rightarrow Y_{0k} = \sigma_\mathcal{R}(r, r_k)$, we derive

$$\begin{aligned} &\mathbb{P}\left(X_{0k} = x_{0k}, R_k = r_k \,|\, S_0 = s_0, R_0 = r\right) = \\ &\quad = \mathbb{P}\left(X_{0k} = x_{0k} \,|\, S_0 = s_0, R_0 = r\right) \\ &\qquad \cdot \mathbb{P}\left(R_k = r_k \,|\, S_0 = s_0, R_0 = r,\ X_{0k} = x_{0k}\right) \\ &\quad \leq \mathbb{P}\left(Y_k = \sigma_\mathcal{R}(r, r_k) \,|\, S_0 = s_0, R_0 = r, X_{0k} = x_{0k}\right). \end{aligned}$$

Finally, we obtain the following counterpart to (4.52):

$$\lambda(r) \leq \prod_{1 \leq k \leq n} \mathbb{P}\left(Y_k = \sigma_{\mathcal{R}}(r, r_k) \,|\, S_0 = s_0, R_0 = r,\ X_{0k} = x_{0k}\right). \tag{4.57}$$

The probabilities in (4.57) now correspond to measures

$$\mu_{Y|(X=x,R=r,S=s)}, \tag{4.58}$$

where $x \in D_{\mathcal{S}}$, $r \in \mathcal{R}$, $s \in \mathcal{S}$. Again, suppose the measures $\eta_{Y|(X=x)}$ $(x \in D_{\mathcal{S}})$ to provide upper bounds:

$$\forall\, x \in S_{\mathcal{S}} \,:\, \eta_{Y|(X=x)} \geq \bigvee_{s \in \mathcal{S}, r \in \mathcal{R}} \mu_{Y|(X=x,R=r,S=s)}.$$

As before, these measures define a generalized similarity profile $G$ which is an upper approximation of the probabilistic similarity profile $H_{\Sigma}$. The approximation of (4.51) is now given by

$$\lambda(r) \leq \prod_{1 \leq k \leq n} \eta_{Y|(X=\sigma_{\mathcal{S}}(s_0,s_k))}(\sigma_{\mathcal{R}}(r, r_k)).$$

**Proposition 4.39.** Consider a CBI problem $\langle \Sigma, s_0 \rangle$ and suppose the generalized hypothesis $G$ to satisfy $G(x) \geq \bigvee_{r \in \mathcal{R}, s \in \mathcal{S}} \mu_{Y|(X=x,R=r,S=s)}$ for all $x \in D_{\mathcal{S}}$. The function

$$\overline{\lambda} : r \mapsto \prod_{\langle s, \varphi(s) \rangle \in \mathcal{M}} [G(\sigma_{\mathcal{S}}(s_0, s))](\sigma_{\mathcal{R}}(r, \varphi(s))) \tag{4.59}$$

is an upper approximation of the likelihood function (4.55), i.e., $\lambda(r) \leq \overline{\lambda}(r)$ for all $r \in \mathcal{R}$. □

When comparing the reasoning schemes based on (4.54) and (4.59) one can see that, on the one hand, the aggregation by means of the minimum operator in (4.54) yields less precise approximations than the aggregation by means of the product operator. On the other hand, however, the measures $G(x)$ employed in (4.59) will generally approximate the classes of underlying probability measures more accurately since these classes are smaller than the classes approximated by the corresponding measures in (4.54).

Interestingly enough, (4.54) and (4.59) define natural generalizations of the constraint-based inference scheme (3.2). Particularly, we recover (3.2) as a special case of (4.54) resp. (4.59) with $G(x) \equiv 1_{D_{\mathcal{S}} \cap [H(x),1]}$. Recall that each value $[G(x)](y)$ is thought of as an upper approximation of the probability that the similarity of two outcomes is $y$ given that the similarity of the respective inputs is $x$. In the context of probabilistic CBI, the constraint-based inference scheme (3.2) can thus be interpreted as an "all or nothing" approach in the sense that a similarity degree $y$ is either regarded as being completely possible ($H(x) \leq y$ resp. $[G(x)](y) = 1$) or as being completely impossible ($y < H(x)$ resp. $[G(x)](y) = 0$).

## 4.9 Summary and remarks

### Summary

- In this chapter, a probabilistic generalization of the constraint-based approach to CBI (cf. Chapter 3) has been proposed. To this end, the basic concepts of a similarity profile and a similarity hypothesis have been replaced by corresponding probabilistic counterparts. A probabilistic formalization of the CBI hypothesis seems particularly suitable since it emphasizes the heuristic nature of this assumption. Again, this formalization guarantees a clear semantics underlying case-based inference.
- Based on the probabilistic model of CBI, the learning of similarity hypotheses can be realized within the context of statistical inference. Corresponding approaches have been developed in Section 4.3.
- Various approaches to probabilistic inference have been proposed. A major problem of probabilistic CBI is the fact that the random variables (similarity relations between cases) involved in the inference process are not stochastically independent. This leads to difficulties when it comes to combining probabilistic evidence concerning the unknown outcome $\varphi(s_0)$ which has been derived from different cases.
- To overcome this problem, a simplifying assumption of independence has been made in Section 4.3.1. On the basis of this assumption, a generalization of the constraint-based inference scheme could be developed which produces a nested sequence of credible output sets with associated confidence levels. Despite the admitted naïvety of the independence assumption, quite satisfactory results have been obtained in the experimental studies (for regression and label ranking problems) in Section 4.4.
- In connection with these studies, it is again worth mentioning that CBI in a sense unifies diverse types of prediction problems, and that it is quite general and widely applicable. In fact, since no kind of transitivity is assumed for the similarity measures and, hence, the structure of the input space $\mathcal{S}$ and the output space $\mathcal{R}$ might be weaker than that of a metric space, CBI is applicable in many situations where standard methods (e.g. from statistics) cannot be used. Besides, CBI combines advantages from both, instance-based and model-based (statistical) learning: As an instance-based approach it requires less structural assumptions than (parametric) statistical methods, and yet it allows for specifying the uncertainty related to predictions.
- The problem of combining different pieces of probabilistic evidence concerning the unknown outcome $\varphi(s_0)$ has been reconsidered in Section 4.5 within the framework of information fusion. A main idea of the inference scheme proposed in this section is to look at previous cases as individual information sources. Loosely speaking, each observed case serves as an uncertain piece of informa-

tion which provides evidence concerning the outcome associated with the new input. This evidence is represented in the form of a belief function. CBI comes down to combining these individual pieces of evidence, i.e., the belief functions derived from individual cases. In Section 4.5.3, a convex combination has been proposed as a reasonable aggregation. Thus, we finally obtain from CBI a characterization of the solution to the new problem in the form of a belief function over the set of possible candidates.

- A convex combination of individual belief functions requires the weighting of the involved cases. In this connection, Section 4.6 has touched on the idea of rating (and eventually discounting) cases in order to optimize predictive performance. The assessment of individual cases can also support the design of optimal memories.
- In addition to the generalizations already introduced in Chapter 3 (profiles supporting the ideas of utilizing a fixed memory and of drawing inferences from the $k$ most similar cases as well as profiles associated with individual inputs) we have proposed similarity profiles of higher order (Section 4.7.1) and $\mathcal{H}$-profiles (Section 4.7.2). Based on these concepts, it is possible to develop inference schemes which avoid the problem of combining individual pieces of probabilistic evidence.
- Similarity hypotheses based on generalized uncertainty measures have been introduced in Section 4.8, where they have been employed in order to realize an approximate probabilistic inference procedure. This approach allows for deriving upper approximations of a likelihood function characterizing the unknown outcome.
- This chapter has revealed interesting relations between CBI on the one side, and probability theory and statistics on the other side. Firstly, realizing CBI in the context of probabilistic reasoning and statistical inference makes a powerful methodological framework accessible to CBI (cf. Section 4.2). Secondly, case-based inference provides the basis of a generalized approach to statistical modeling (cf. Section 4.3) and can be seen as a step toward an extended (probabilistic) approach to experience-based reasoning.

### Remarks

- It should be stressed that the probabilistic formalization developed in this chapter does not rely on very specific assumptions but emerges quite naturally as a generalization of the constraint-based approach in connection with the probabilistic modeling of the occurrence of inputs. Moreover, this formalization is not related to particular inference schemes. Rather, it provides the basic concepts for "translating" a CBI problem into one of probabilistic reasoning and, hence, for deriving such schemes.

- Probabilistic models, particularly Bayesian networks, have been used in connection with case-based reasoning by several authors [8, 54, 63, 275, 317]. As main differences between most of these approaches and our work let us emphasize two points. Firstly, the concept of similarity is often *derived* from that of probability or vice versa. In [317], for instance, similarity is *interpreted* as a certain probability related to a classification task. In our approach, similarity is seen as an additional and independent concept, containing important information which is not already (implicitely) provided by other concepts: A similarity hypothesis (including the function $h$ and the respective measures $\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$) *complements* an original model description and, hence, supports the (heuristic) derivation of inferences about a system.

  Secondly, a probabilistic model is often related to the features (attributes) of cases (at the system level) *directly*, whereas our formalization proceeds from the similarity level. For example, the problem considered in [317] is to estimate the (conditional) probability $\mathbb{P}(C_j \,|\, c)$ of a case $c = (f_1, \ldots, f_n)$ to belong to category $C_j$ as a function of the features $f_k$. Thus, whereas inference results are here derived from cases at the system level, our approach derives them from the *similarity structures* at the similarity level.
- The approaches in Section 4.3 and Section 4.5 are closely related to algorithms based on the Nearest Neighbor principle, such as instance-based learning. However, there are also important differences. Without going into detail, let us briefly mention two points. Firstly, CBI fits an explicit model of the CBI hypothesis to the current application, whereas IBL uses this assumption by more indirect means. Secondly, IBL derives point-estimations in the form of, say, numeric outputs or predicted class labels.[32] As opposed to this, our approach provides a characterization of the unknown outcome by means of credible outputs sets or, more generally, an uncertainty measure. This way, it allows for taking the validity of the CBI hypothesis into account. In fact, a prediction will hardly support specific outcomes by means of high degrees of plausibility if the similarity structure of the system under consideration is poorly developed, thus indicating that the NN principle should be applied with caution.
- Needless to say, hinting at the credibility of proposed solutions seems indispensable for certain applications of CBI, such as evidential reasoning in medicine [35]. In this connection, one has to bear in mind that an indication of the confidence in a prediction is possible only under certain conditions. Firstly, the quantification of the *deviation* of the prediction from the true outcome assumes the set of outputs to be equipped with a concept of *distance* or *similarity*.[33] Secondly, quantifying the possibility of an *error* (an incorrect prediction) or, more generally, a certain deviation from the true outcome requires the instance

[32] Yet it should be mentioned that the (distance-weighted) relative frequency of a class label among the $k$ nearest neighbors is often interpreted as a (posterior) probability of that class. Such an interpretation, however, seems to be reasonable only under special assumptions.

[33] Observe that (instance-based) concept learning algorithms do generally not assume a similarity or distance measure over the set of classes (= outcomes).

space to be equipped with some *probabilistic structure*. In fact, the basic question which has to be answered concerns the probability that the observations have emerged from a data-generating process which would suggest a different prediction. Next to standard probabilistic models (in the form of probability distributions), for example, so-called *random fields* over the instance space, i.e., some $n$-dimensional Euclidean space, have been used in NN classification [254]. Of course, the definition (and justification) of a probabilistic structure becomes more involved if the instance space is not given as a simple numeric space, which is rarely the case in CBR. We avoid this problem by defining a corresponding structure (namely a PSP) over the (numeric) similarity space.

- In [142], it is shown that a special version of the CBR hypothesis is correct *on average*, in the sense that problems with similar features are more likely to have the same solution, given that the similarity measure is appropriately defined. Two important differences to our approach deserve mentioning: Firstly, we assume a similarity measure to be given, that is, our approach does not require the specification of an ideal measure but remains valid regardless of the similarity measure employed. Secondly, we are not directly concerned with the probability of a correct versus incorrect prediction (which only makes sense if the output set is small, a requirement rarely fulfilled in CBR), but rather with the derivation of credible sets which are likely to cover the true output (solution).
- Regarding the aspect of uncertainty in CBR, the importance of being able to assign degrees of confidence to predictions has been pointed out by several authors (e.g. [65]). In [66], different confidence measures for case-based (nearest neighbor) predictions are proposed and evaluated, and this work has been continued in [81] in connection with a concrete CBR application (Spam filtering). More generally, the problem to characterize the reliability of an estimation has recently received attention in the machine learning field as well [243, 301]. Again, however, note that assessing the confidence of a single point estimate, as done in the aforementioned papers, is quite different from our goal to derive predictions in the form of credible sets.
- A generalization of the $k$-Nearest Neighbor classifier which is closely related to our approach of Section 4.5.3 has been developed in [84]. In this method, each neighbor $x_\imath$ of a new query point (pattern) $x_0$ specifies the unknown class $c_0 \in C$ of $x_0$ by means of a belief function $\mathsf{Bel}_\imath$ resp. an associated mass distribution $\mathsf{m}_\imath$ such that
$$\mathsf{m}_\imath(\{c_\imath\}) = \alpha_\imath, \quad \mathsf{m}_\imath(C) = 1 - \alpha_\imath. \tag{4.60}$$
The weight $0 < \alpha_\imath < 1$ expresses the degree of support of the hypothesis $c_0 = c_\imath$. A further generalization, in which the class of a training pattern is specified by means of a possibility distribution on $C$ (rather than by a precise class label), has been proposed in [85].

As in our approach, individual predictions are hence specified in terms of belief functions. Still, let us mention some major differences. Firstly, (4.60) assigns a positive mass only to the class $c_\imath$ (and to the complete set of classes $C$). Thus, an observed case (pattern) can either support its own class – as in the original NN method – or it can express ignorance by attaching a (more or less large) mass to $C$. As opposed to this, a case can also support *similar* outcomes in our approach. In fact, it might even support outcomes which are quite different from its own output. In this connection, it deserves mentioning that [84] does not assume a similarity structure over the set $C$ of classes.

Secondly, the masses themselves are derived from an additional concept in our approach, namely from a model in the form of a probabilistic similarity profile. In (4.60), the degree of support $\alpha_\imath$ is assumed to be a decreasing function of the distance between $x_0$ and $x_\imath$.[34]

A further difference concerns the way of aggregating the predictions induced by different cases. In [84], this is accomplished by means of DEMPSTER's rule of combination. As discussed in Section 4.5.3, we advocate a different aggregation method – namely a convex combination – since we do not consider the belief functions provided by neighbored cases as distinct pieces of evidence.

- As has been pointed out in Section 4.6, the assessment of cases can support the selection of useful cases to be memorized. Of course, further aspects, including the size, density, and distribution of a case base, have to be taken into account for maintaining an optimal memory of cases [253, 282]. In [356], an interesting model of the *competence* of a case base has been proposed which is based on the concept of so-called *competence groups*. However, by assuming regular problem spaces and cases which are representative of the target problem, this model disregards the problem of exceptionality of cases completely. Thus, it takes a more global look at a memory, whereas we have concentrated on properties of individual cases. Needless to say, these two aspects could reasonably complement each other.
- Interestingly enough, the probabilistic nature of analogical reasoning or, more generally, inductive generalization has been emphasized by philosophically minded scholars for a long time. In fact, the traditional approach to deciding whether an analogy is reasonable, which goes back at least to MILL [266], is to consider each observed similarity as a piece of extra evidence in favor of the correctness of the conclusion. MILL's idea is to associate a “probability” with an analogical inference pattern such as the following:

$$F(a) \wedge \mathrm{sim}(a, b) \Rightarrow F(b).$$

[34] See [421] for an approach to adapting this function in an optimal way.

(If predicate $F$ applies to object $a$ and if $a$ is similar to $b$, then $F$ does also apply to $b$.) According to his probabilistic interpretation, the strength (probability of correctness) of the above rule corresponds to the similarity between the objects $a$ and $b$. Thus, similarity provides a probabilistic basis of inference.

# 5. Fuzzy Set-Based Modeling of Case-Based Inference I

A close connection between fuzzy set-based (approximate reasoning) methods and the inference principle underlying similarity-based (case-based) reasoning has been pointed out recently [99, 407]. Besides, some attempts at combining case-based reasoning (or, more generally, analogical reasoning) and methods from fuzzy set theory have already been made [408], including the use of fuzzy sets for supporting the computation of similarities of situations in analogical reasoning [144], the formalization of aspects of analogical reasoning by means of similarity relations between fuzzy sets [48], the use of fuzzy set theory in case indexing and retrieval [209, 214], the case-based learning of fuzzy concepts from fuzzy examples [295], the use of fuzzy predicates in the derivation of similarities [40], and the integration of case-based and rule-based reasoning [138]. See [45, 49] for a more general framework of analogical reasoning.

This chapter continues this promising line of research. It is argued that fuzzy rules in conjunction with associated inference procedures provide a convenient framework for modeling the CBI hypothesis and for supporting the task of case-based inference as outlined in Section 2.4.

The remaining part of the chapter is organized as follows: Even though we assume the reader to be familiar with basics of fuzzy set theory, we recall the most important concepts from possibility theory in Section 5.1. The basic CBI framework we proceed from and the key idea of fuzzy rule-based modeling of the CBI hypothesis are introduced in Section 5.2. Diverse types of extensions of the basic model will then be discussed in Sections 5.3 and 5.4. Section 5.5 presents some experimental studies in the field of classification. The idea of calibrating a CBI model by combining qualitative modeling techniques with data-driven optimization methods is addressed in Section 5.6. Finally, some connections between the approach introduced in this chapter and related approaches in the field of fuzzy set theory are discussed in Section 5.7.

## 5.1 Background on possibility theory

In this section, we recall some basic concepts from possibility theory, as far as required for the current chapter. Possibility theory deals with “degrees of possibility”. The term “possibility” is hence employed as a *graded* notion, much in the same way as the term “probability”. At first sight, this might strike as odd since “possibility” is usually considered a two-valued concept in natural language (something is possible or not). Before turning to more technical aspects, let us therefore make some brief remarks on the semantics underlying the notion of “possibility” as used in possibility theory.

Just as the concept of probability, the notion of possibility can have different semantic meanings. To begin with, it can be used in the (physical) sense of a “degree of ease”. One might say, for instance, that it is more possible for Hans to have two eggs for breakfast than eight eggs, simply because eating two eggs is more easy (feasible, practicable) than eating eight eggs [416]. However, as concerns the use in most applications, and in this book in particular, possibility theory is considered as a means for representing uncertain knowledge, that is, for characterizing the epistemic state of an agent. For instance, given the information that Hans has eaten *many* eggs, one is clearly uncertain about the precise number. Still, three eggs appears somewhat more plausible (possible) than two eggs, since three is more compatible with the linguistic quantifier “many” than two.

It is important to note that a degree of possibility, as opposed to a degree of probability, is not necessarily a number. In fact, for many applications it is sufficient, and often even more suitable, to assume a qualitative (ordinal) scale with possibility degrees ranging from, e.g., “not at all” and “hardly” to “fairly” and “completely” [251, 127]. Still, possibility degrees can also be measured on the cardinal scale $[0, 1]$, again with different semantic interpretations. For example, possibility theory can be related to probability theory, in which case a possibility degree can specify, e.g., an upper probability bound [122]. For convenience, possibility degrees are often coded by numbers from the unit interval even within the qualitative framework of possibility theory.

As a means of representing uncertain knowledge, possibility theory makes a distinction between the concepts of *certainty* and *plausibility* of an event. As opposed to probability theory, possibility theory does not claim that the confidence in an event is determined by the confidence in the complement of that event and, consequently, involves non-additive measures of uncertainty. Taking the existence of two quite opposite but complementary types of knowledge representation and information processing into account, two different versions of possibility theory will be outlined in the following. For a closer discussion refer to [131] and [104].

### 5.1.1 Possibility distributions as generalized constraints

A key idea of possibility theory as originally introduced by Zadeh [416] is to consider a piece of knowledge as a (generalized) constraint that excludes some “world states” (to some extent). Let $\Omega$ be a set of worlds conceivable by an agent, including the “true world” $\omega_0$. With (incomplete) knowledge $\mathcal{K}$ about the true world one can then associate a possibility measure $\Pi_{\mathcal{K}}$ such that $\Pi_{\mathcal{K}}(A)$ measures the compatibility of $\mathcal{K}$ with the event (set of worlds) $A \subseteq \Omega$, i.e., with the proposition that $\omega_0 \in A$. Particularly, $\mathit{\Pi}_{\mathcal{K}}(A)$ becomes small if $\mathcal{K}$ excludes each world $\omega \in A$ and large if at least one of the worlds $\omega \in A$ is compatible with $\mathcal{K}$. More specifically, the finding that $\mathcal{A}$ is incompatible with $\mathcal{K}$ to some degree corresponds to a statement of the form $\mathit{\Pi}_{\mathcal{K}}(A) \leq p$, where $p$ is a possibility degree taken from an underlying possibility scale $P$.

The basic informational principle underlying the possibilistic approach to knowledge representation and reasoning is stated as a *principle of minimal specificity*:[1] In order to avoid any unjustified conclusions, one should represent a piece of knowledge $\mathcal{K}$ by the *largest* possibility measure among those measures compatible with $\mathcal{K}$, which means that the inequality above is turned into an equality: $\mathit{\Pi}_{\mathcal{K}}(A) = p$. Particularly, complete ignorance should be modeled by the measure $\mathit{\Pi} \equiv 1$.

Knowledge $\mathcal{K}$ is usually expressed in terms of a *possibility distribution* $\pi_{\mathcal{K}}$, a $\mathit{\Omega} \longrightarrow P$ mapping related to the associated measure $\mathit{\Pi}_{\mathcal{K}}$ through

$$\mathit{\Pi}_{\mathcal{K}}(A) = \sup_{\omega \in A} \pi_{\mathcal{K}}(\omega).$$

Thus, $\pi_{\mathcal{K}}(\omega)$ is the degree to which world $\omega$ is compatible with $\mathcal{K}$.

Apart from the boundary conditions $\mathit{\Pi}_{\mathcal{K}}(\mathit{\Omega}) = 1$ (at least one world is fully possible) and $\mathit{\Pi}_{\mathcal{K}}(\emptyset) = 0$, the basic axiom underlying possibility theory after ZADEH involves the maximum-operator:

$$\mathit{\Pi}_{\mathcal{K}}(A \cup B) = \max\left\{\mathit{\Pi}_{\mathcal{K}}(A),\, \mathit{\Pi}_{\mathcal{K}}(B)\right\}. \tag{5.1}$$

In plain words, the possibility (or, more precisely, the upper possibility-bound) of the union of two events $A$ and $B$ is the maximum of the respective possibilities (possibility-bounds) of the individual events.

As constraints are naturally combined in a conjunctive way, the possibility measures associated with two pieces of knowledge, $\mathcal{K}_1$ and $\mathcal{K}_2$, are combined by using the minimum-operator:

$$\pi_{\mathcal{K}_1 \wedge \mathcal{K}_2}(A) = \min\{\pi_{\mathcal{K}_1}(A), \pi_{\mathcal{K}_2}(A)\}$$

for all $A \subseteq \mathit{\Omega}$. Note that $\pi_{\mathcal{K}_1 \wedge \mathcal{K}_2}(\mathit{\Omega}) < 1$ indicates that $\mathcal{K}_1$ and $\mathcal{K}_2$ are not fully compatible, i.e., that $\mathcal{K}_1 \wedge \mathcal{K}_2$ is contradictory to some extent.

The distinction between possibility and certainty of an event is reflected by the existence of a so-called *necessity measure* $\mathcal{N}_{\mathcal{K}}$ that is dual to the possibility measure $\mathit{\Pi}_{\mathcal{K}}$. More precisely, the relation between these two measures is given by

$$\mathcal{N}_{\mathcal{K}}(A) = 1 - \mathit{\Pi}_{\mathcal{K}}(\mathit{\Omega} \setminus A) \tag{5.2}$$

for all $A \subseteq \mathit{\Omega}$:[2] An event $A$ is necessary in so far as its complement (logical negation) is not possible.

[1] This principle plays a role quite comparable to the maximum entropy principle in probability theory.

[2] If the possibility scale $P$ is not the unit interval $[0, 1]$, the mapping $1 - (\cdot)$ on the right-hand side of (5.2) is replaced by an order-reversing mapping of $P$.

Worth mentioning is the close relationship between possibility theory and fuzzy sets. In fact, the idea of Zadeh [416] was to induce a possibility distribution from knowledge stated in the form of vague linguistic information and represented by a fuzzy set. Formally, he postulated that $\pi_{\mathcal{K}}(\omega) = \mu_F(\omega)$, where $\mu_F$ is the membership function of a fuzzy set $F$. To emphasize that $\omega$ plays different roles on the two sides of the equality, the latter might be written more explicitly as $\pi_{\mathcal{K}}(\omega \,|\, F) = \mu(F \,|\, \omega)$: Given the knowledge $\mathcal{K}$ that $\omega$ is an element of the fuzzy set $F$, the possibility that $\omega_0 = \omega$ is evaluated by the degree to which the fuzzy concept (modeled by) $F$ is satisfied by $\omega$. To illustrate, suppose that world states are simply integer numbers. The uncertainty related to the vague statement that "$\omega_0$ is a small integer" ($\omega_0$ is an element of the fuzzy set $F$ of small integers) might be translated into a possibility distribution that lets $\omega_0 = 1$ appear fully plausible ($\mu_F(1) = 1$), whereas, say, 5 is regarded as only more or less plausible ($\mu_F(5) = 1/2$) and 10 as impossible ($\mu_F(10) = 0$).

### 5.1.2 Possibility as evidential support

Possibility theory as outlined above provides the basis of a generalized approach to constraint propagation, where constraints are expressed in terms of possibility distributions (fuzzy sets) rather than ordinary sets (which correspond to the special case of $\{0, 1\}$-valued possibility measures). A constraint usually corresponds to a piece of knowledge that excludes certain alternatives as being impossible (to some extent). This "knowledge-driven" view of reasoning is complemented by a, say, "data-driven" view that leads to a different type of possibilistic calculus. According to this view, the statement that "$\omega$ is possible" is not intended to mean that $\omega$ is provisionally accepted in the sense of not being excluded by some constraining piece of information, but rather that $\omega$ is indeed supported or, say, confirmed by already observed facts (in the form of examples or data).

To distinguish the two meanings of a possibility degree, we shall denote a degree of *evidential support* or *confirmation* of $\omega$ by $\delta(\omega)$,[3] whereas $\pi(\omega)$ denotes a degree of compatibility.

To illustrate, suppose that the values a variable $V$ can assume are a subset of $\mathcal{V} = \{1, 2, \ldots, 10\}$ and that we are interested in inferring which values are possible and which are not. In agreement with the example-based (data-oriented) view, we have $\delta(v) = 1$ as soon as the instantiation $V = v$ has indeed been observed and $\delta(v) = 0$ otherwise. The knowledge-driven approach can actually not exploit such examples, since an observation $V = v$ does not exclude the possibility that $V$ can also assume any other value $v' \neq v$. As can be seen, the data-driven and the knowledge-driven approach are intended, respectively, for expressing *positive* and *negative* evidence [108]. As examples do express positive evidence, they do never change the distribution $\pi \equiv 1$. This distribution would only be changed if

[3] In [393], this type of distribution is called $\sigma$-distribution.

we *knew* from some other information source, e.g., that $V$ can only take values $v \geq 6$, in which case $\pi(v) = 1$ for $v \geq 6$ and $\pi(v) = 0$ for $v \leq 5$.

The difference between modeling positive and negative evidence becomes especially clear when it comes to expressing complete ignorance. As already mentioned above, this situation is adequately captured by the possibility distribution $\pi \equiv 1$: If nothing is known, there is no reason to exclude any of the worlds $\omega$, hence each of them remains completely possible. At the same time, complete ignorance is modeled by the distribution $\delta \equiv 0$. The latter does simply express that none of the worlds $\omega$ is actually supported by observed data.

Within the context of modeling evidential support, possibilistic reasoning accompanies a process of data accumulation. Each observed fact, $\phi$, guarantees a certain degree of possibility of some world state $\omega$, as expressed by an inequality of the form $\delta_\phi(\omega) \geq d$. The basic informational principle is now a principle of *maximal informativeness* that suggests adopting the smallest distribution among those compatible with the given data and, hence, to turn the above inequality into an equality. The accumulation of observations $\phi_1$ and $\phi_2$ is realized by deriving a distribution that is pointwise defined by

$$\delta_{\phi_1 \wedge \phi_2}(\omega) = \max\{\delta_{\phi_1}(\omega), \delta_{\phi_2}(\omega)\}.$$

As can be seen, adding new information has quite an opposite effect in connection with the two types of possibilistic reasoning: In connection with the knowledge-driven or constraint-based approach, a new constraint can only reduce possibility degrees, which means turning the current distribution $\pi$ into a smaller distribution $\pi' \leq \pi$. In connection with the data-driven or example-based approach, new data can only increase (lower bounds to) degrees of possibility.

Closely related to the view of possibility as evidential support is a set-function that was introduced in [121], called measure of "guaranteed possibility": $\Delta(A)$ is the degree to which *all* worlds $\omega \in A$ are possible, whereas an event $A$ is possible in the sense of the usual measure of "potential possibility", namely $\Pi(A)$ as discussed above, if at least one $\omega \in A$ is possible.[4] For the measure $\Delta$, the characteristic property (5.1) becomes

$$\Delta(A \cup B) = \min\{\Delta(A), \Delta(B)\}.$$

## 5.2 Fuzzy rule-based modeling of the CBI hypothesis

Rule-based modeling plays an important role in fuzzy systems research and will also turn out to be useful in the context of case-based inference. Fuzzy rules provide a local, rough and soft specification of the relation between variables $X$

[4] The latter semantics is clearly in line with the measure-theoretic approach underlying probability theory.

and $Y$ ranging on domains $D_X$ and $D_Y$, respectively [124]. They are generally expressed in the form "if $X$ is $A$ then $Y$ is $B$," where $A$ and $B$ are fuzzy sets associated with symbolic labels and modeled by means of membership functions on $D_X$ resp. $D_Y$.[5]

There are several aspects which motivate the use of fuzzy rules in connection with case-based reasoning [100, 205]. Firstly, the CBI hypothesis itself corresponds to an *if-then* rule: "If two inputs are similar, then the associated outcomes are similar as well." Secondly, the notion of *similarity*, which lies at the heart of case-based reasoning, is also strongly related to the theory of fuzzy sets. Indeed, one of the main interpretations of the membership function of a fuzzy set is that of a similarity relation, i.e., degrees of membership can be thought of as degrees of similarity [126]. Thirdly, linked with the framework of possibility theory, fuzzy sets provide a tool for the modeling and processing of *uncertainty*. In connection with the *heuristic* character of CBR, this aspect seems to be of special importance. As already mentioned in Chapter 1, the CBI principle should not be understood as a deterministic rule. Within the context of fuzzy rules considered in this chapter, it will rather be interpreted in the following sense: "If two inputs are similar, it is *possible* that the associated outcomes are similar as well."

At a formal level, fuzzy rules can be modeled as possibility distributions constrained by a combination of the membership functions which define the antecedent and consequent part of the rule, where the concrete form of the constraint depends on the interpretation of the rule [124]. This way, they relate the concepts of *similarity* and *uncertainty*, thus providing the basis for methods of uncertain similarity-based inference. This is the main reason for their convenience as formal models of the CBI hypothesis

### 5.2.1 Possibility rules

The aforementioned interpretation of the CBI hypothesis is nicely captured by means of a so-called *possibility rule*, a special type of conjunction-based fuzzy rule. A possibility rule involving fuzzy sets $A$ and $B$, subsequently symbolized by $A \rightharpoondown B$, corresponds to the statement that "the more $X$ is $A$, the more *possibly* $B$ is a range for $Y$." More precisely, it can be interpreted as a collection of rules "if $X = x$, it is possible at least to the degree $A(x)$ that $B$ is a range for $Y$." The intended meaning of this kind of *possibility-qualifying* rule is captured by the following constraint which guarantees a certain lower bound to the possibility $\delta(x, y)$ that the tuple $(x, y)$ is an admissible instantiation of the variables $(X, Y)$:

$$\delta(x, y) \geq \min\{\, A(x),\, B(y) \,\}. \tag{5.3}$$

As suggested by the rule-based modeling of the relation between $X$ and $Y$, these variables often play the role of an input and an output, respectively, and one

[5] We shall usually use the same notation for a label, the name of an associated fuzzy set, and the membership function of this set. Thus, $A(x)$ is the degree of membership of the element $x$ in the fuzzy set $A$.

is interested in possible values of $Y$ while $X$ is assumed to be given. By letting $\delta(y \,|\, x) \overset{\text{df}}{=} \delta(x,y)$, the constraint (5.3) can also be considered as a lower bound to a *conditional* possibility distribution. That is, given the value $X = x$, the possibility that $Y = y$ is lower-bounded by $\delta(x,y)$ according to (5.3). Observe that *nothing* is said about $Y$ in the case where $A(x) = 0$ since we then obtain the trivial constraint $\pi(y \,|\, x) \geq 0$. Besides, it should be noticed that the lower bound-interpretation is also consistent with conditional distributions $\delta(\cdot \,|\, x)$ which are not normalized, i.e., for which $\sup_y \delta(y \,|\, x) < 1$ (cf. Section 5.1).

### 5.2.2 Modeling the CBI hypothesis

The basic framework we shall proceed from in this chapter is a special type of generalized non-deterministic CBI setup (see Definition 2.7 and Remark 2.8 in Section 2.4.2). As in Chapters 3 and 4, a case $c$ is a tuple $\langle s, r \rangle \in \mathcal{C} = \mathcal{S} \times \mathcal{R}$ consisting of an input $s \in \mathcal{S}$ and an associated output $r \in \mathcal{R}$. However, we do no longer assume that an input determines a unique outcome, i.e., cases $c = \langle s, r \rangle$ and $c' = \langle s', r' \rangle$ such that $s = s'$ but $r \neq r'$ might be encountered. In fact, the assumption of a functional relation $\varphi : \mathcal{S} \longrightarrow \mathcal{R}$ mapping inputs to unique outcomes would be too restrictive for the type of applications we have in mind in connection with the possibilistic approach. Rather, $\varphi$ is now defined as a relation

$$\varphi \subseteq \mathcal{S} \times \mathcal{R} \tag{5.4}$$

and corresponds to a set of potential observations, i.e., existing (but perhaps not yet encountered) cases. As before, we assume data to be given in the form of a memory

$$\mathcal{M} = \big\{ \langle s_1, r_1 \rangle, \langle s_2, r_2 \rangle, \ldots, \langle s_n, r_n \rangle \big\}$$

of observed cases. As an aside, note that $\mathcal{M}$ was formally treated as a sequence rather than a set in Chapters 3 and 4. This is not necessary within the possibilistic framework of this section. Moreover, we can abandon the assumption that $\mathcal{S}$ and $\mathcal{R}$ are countable.

As before, our focus is on case-based inference: Given a new input $s_0 \in \mathcal{S}$, the task is to predict the outcome $r_0 \in \mathcal{R}$ associated with $s_0$. This actually comes down to predicting the set $\{ r \in \mathcal{R} \,|\, \langle s_0, r \rangle \in \varphi \}$ of potential outcomes, since we do no longer assume uniqueness. To this end, we shall derive a quantification of the *possibility* that $r_0 = r$, i.e., $\langle s_0, r \rangle \in \varphi$, for each outcome $r \in \mathcal{R}$. As will be seen in the remainder of this chapter, this kind of prediction makes the formulation of rather general types of queries possible, especially if $s_0$ is allowed to be incompletely specified.

The basic idea of the approach discussed in this chapter is to use a possibility rule as defined above in order to formalize the CBI hypothesis. In fact, interpreting the variables $X$ and $Y$ as degrees of similarity between two inputs and two outputs, respectively, and $A$ and $B$ as fuzzy sets of "large similarity degrees" (with strictly increasing membership functions) amounts to expressing the following version of the CBI hypothesis: "The more similar two inputs are, the more *possible* it is that the corresponding outcomes are similar" [99]. In the same way as the probabilistic model of Chapter 4, this formalization takes the heuristic nature of the CBI hypothesis into account. In fact, it does not impose a deterministic constraint, but only concludes on the *possibility* of the outcomes to be similar.

In the sense of the above principle, an observed case $\langle s_1, r_1 \rangle \in \mathcal{M}$ is taken as a piece of evidence which qualifies similar (hypothetical) cases $\langle s, r \rangle$ as being possible. According to (5.3) it induces lower bounds[6]

$$\delta(s,r) \geq \min \{ \sigma_{\mathcal{S}}(s,s_1), \, \sigma_{\mathcal{R}}(r,r_1) \} \tag{5.5}$$

to the possibility that $\langle s, r \rangle \in \varphi$. This can be interpreted as a similarity-based *extrapolation* of case-based information: The observation $\langle s_1, r_1 \rangle$ is considered as a typical case or, say, prototype, which is extrapolated in accordance with the CBI hypothesis. The more similar $\langle s, r \rangle$ and $\langle s_1, r_1 \rangle$ are in the sense of the (joint) similarity measure

$$\sigma_{\mathcal{C}} : \big(\langle s,r \rangle, \langle s',r' \rangle\big) \mapsto \min \big\{\sigma_{\mathcal{S}}(s,s'), \sigma_{\mathcal{R}}(r,r')\big\} \, , \tag{5.6}$$

the more plausible becomes the (hypothetical) case $\langle s, r \rangle$ and, hence, the larger is the (lower) possibility bound (5.5). In other words, a high degree of possibility is assigned to a hypothetical case as soon as the *existence* of a very similar case is guaranteed (by observation).

Applying (5.5) to all cases in the memory $\mathcal{M}$ we obtain the possibility distribution $\delta_{\mathcal{C}}$ defined by

$$\delta_{\mathcal{C}}(s,r) = \max_{1 \leq \imath \leq n} \min \big\{\sigma_{\mathcal{S}}(s,s_\imath), \sigma_{\mathcal{R}}(r,r_\imath)\big\} \tag{5.7}$$

for all $c = \langle s, r \rangle \in \mathcal{S} \times \mathcal{R}$. This distribution can be interpreted as a possibilistic approximation of the relation $\varphi$ in (5.4). It is of provisional nature and actually represents lower bounds to possibility degrees (the equality in (5.7) is justified by the principle of *maximal informativeness*, see Section 5.1.2). In fact, the degree of possibility assigned to a case $c$ may increase when gathering further evidence by observing new sample cases, as reflected by the application of the maximum operator in (5.7).

Observe that similarity degrees (on the right-hand side) are turned into possibility degrees (on the left-hand side) by virtue of the functional relation (5.7). In fact, the latter reveals at a formal level that – according to our formalization – similarity is in direct correspondence with possibility: From the similarity of a case $\langle s, r \rangle$ to an observed case, (5.7) concludes on the possibility of this case itself.

The distribution (5.7) can be taken as a point of departure for various inference tasks. In particular, given a new input $s_0$, a prediction of the associated outcome $r_0$ is obtained in the form of the conditional distribution $\delta_{s_0}$ defined by

[6] Without loss of generality, we assume the membership functions of the fuzzy sets of "large similarity degrees" to be given by the identical function id : $x \mapsto x$ on $[0, 1]$.

$$\delta_{s_0}(r) = \delta(r \,|\, s_0) \overset{\mathrm{df}}{=} \max_{1 \leq \imath \leq n} \min \left\{ \sigma_{\mathcal{S}}(s_0, s_\imath),\, \sigma_{\mathcal{R}}(r, r_\imath) \right\}, \tag{5.8}$$

for all $r \in \mathcal{R}$, where $\delta_{s_0}(r)$ denotes the (estimated) *possibility* of the output $r$, i.e., the possibility that $r$ corresponds to the true outcome $r_0$.

EXAMPLE 5.1. The (real-world) AUTOMOBILE DATABASE[7] contains 205 cars, each of which is characterized by 26 attributes. Thus, let a case correspond to a car which is characterized by means of an attribute–value representation including properties, such as its horsepower and fuel-type. For the sake of simplicity, we shall consider only some of the attributes available, i.e., the memory $\mathcal{M}$ is actually a projection of the complete database. One of the attributes, namely the price of a car, has been chosen as the outcome associated with a case. The latter is hence a tuple $\langle s, r \rangle$, where the input $s = (a_1, \ldots, a_L)$ is a vector of attribute values describing a car, and $r$ is the associated price. The similarity between two cars $s$ and $s'$ is defined as a combination of the similarities between the respective attribute values $a_\jmath$ and $a'_\jmath$ $(1 \leq \jmath \leq L)$.

To illustrate, suppose a car to be characterized by only one attribute, namely its horsepower. Thus, the CBI hypothesis should simply be understood in the sense that "cars with similar horsepowers (possibly) have similar prices." Let $\sigma_{\mathcal{S}}(s, s') = \sigma_{hp}(s, s') = \max\{1 - |s - s'|/100, 0\}$. Likewise, let the similarity between two outcomes (= prices) be given by $\sigma_{\mathcal{R}}(r, r') = \max\{1 - |r - r'|/10000, 0\}$. Fig. 5.1 shows the prediction (5.8) for $s_0 = 100$. This prediction corresponds to the "more or less" possible range of prices for the class of cars whose horsepower is 100. As can be seen, the evidence contained in the memory $\mathcal{M}$ of cases strongly supports prices between $\$10,000$ and $\$17,000$. At the same time, however, it does not completely rule out prices which are slightly lower or higher. □

**The possibility distribution $\boldsymbol{\delta_{s_0}}$.** According to (5.8), $r$ is regarded as a possible output if there is a case $\langle s_\imath, r_\imath \rangle$ such that both, $s_\imath$ is close to $s_0$ and $r_\imath$ is close to $r$. Or, if we define the *joint similarity* between the case $\langle s_\imath, r_\imath \rangle$ and the (hypothetical) case $\langle s_0, r \rangle$ according to (5.6), this can be expressed by saying that the case $\langle s_0, r \rangle$ is regarded as possible if the existence of a similar case $\langle s_\imath, r_\imath \rangle$ is confirmed by observation. In other words, a similar case provides evidence for the existence of $\langle s_0, r \rangle$ in the sense of *possibility qualification*.[8]

Following the notational convention of Section 5.1, possibility degrees $\delta_{s_0}(r)$ denote degrees of "guaranteed possibility". Thus, they are actually not considered as degrees of plausibility in the usual sense but rather as degrees of *confirmation* as introduced in Section 5.1.2. More specifically, the distribution $\delta_{s_0} : \mathcal{R} \longrightarrow [0, 1]$ is thought of as a *lower* rather than an upper bound. Particularly, $\delta_{s_0}(r) = 0$ must

[7] Available at `http://www.ics.uci.edu/~mlearn`.

[8] The idea of possibility qualification, already mentioned in Section 5.1, is usually considered in connection with natural language propositions [328, 417]. Here, possibility qualification is casuistic rather than linguistic.

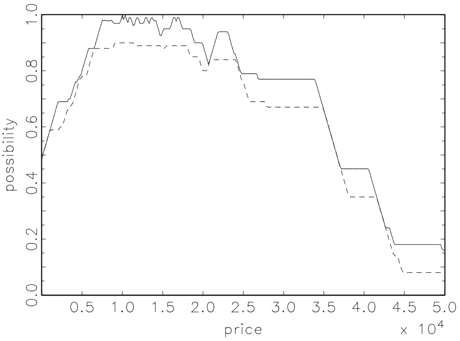


**Fig. 5.1.** Prediction (5.8) of the price of a car with horsepower $s_0 = 100$ (solid line) and prediction (5.32) for $90 \leq s \leq 110$.

not be equated with the impossibility of $r_0 = r$ but merely means that no evidence supporting the outcome $r$ is available so far! In fact, $\delta_{s_0}$ is of provisional nature, and the degree of possibility assigned to an outcome $r$ may increase when gathering further evidence by observing new cases, as reflected by the application of the maximum operator in (5.8). These remarks also make clear that the distribution $\delta_{s_0}$ is not necessarily normalized (in the sense that $\sup_r \delta_{s_0}(r) = 1$). In this connection, also note that there is not necessarily a unique actual world in the sense of the possible worlds semantics [51]. Since $s_0$ is not assumed to have a unique output, $\delta_{s_0}$ rather provides information about the set $\{r \in \mathcal{R} \,|\, \langle s_0, r\rangle \in \varphi\}$ of potential outcomes. Thus, the state of "complete knowledge" corresponds to the distribution $\delta_{s_0}$ with $\delta_{s_0}(r) = 1$ if $\langle s_0, r\rangle \in \varphi$ and $\delta_{s_0}(r) = 0$ otherwise.

In a classification context, where the outcomes $r$ are class labels (i.e., $\mathcal{R}$ is a finite number of classes), the set of all inputs $s \in \mathcal{S}$ with the same output is sometime referred to as a *concept.* When being applied to all $s \in \mathcal{S}$, (5.8) yields "fuzzy" concept descriptions, that is possibilistic approximations of the concepts $C_r$ $(r \in \mathcal{R})$:

$$C_r^{est} = \{(s, \delta_s(r)) \,|\, s \in \mathcal{S}\}, \tag{5.9}$$

where $\delta_s(r)$ is the degree of membership of $s \in \mathcal{S}$ in the fuzzy concept $C_r^{est}$, i.e., $C_r^{est}(s) = \delta_s(r)$. Note that these fuzzy concepts can overlap in the sense that $\min\{C_r^{est}(s), C_{r'}^{est}(s)\} > 0$ for $r \neq r'$ and $s \in \mathcal{S}$ ($s$ has a positive degree of membership in two concepts $C_r^{est}$ and $C_{r'}^{est}$, $r \neq r'$).[9]

**The similarity measures $\boldsymbol{\sigma_{\mathcal{S}}}$ and $\boldsymbol{\sigma_{\mathcal{R}}}$.** Let us make some remarks on the similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$. As mentioned previously, according to (5.8), the

[9] In practice, fuzzy and/or overlapping concepts seem to be the rule rather than the exception [3].

*similarity* of cases is in direct correspondence with the *possibility* assigned to an outcome. Roughly speaking, the principle expressed by (the fuzzy rule underlying) equation (5.8) gives rise to turn similarity into possibilistic support. Consequently, $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ are thought of as, say, support measures rather than similarity measures in the usual sense. They do actually serve the same purpose as the weight functions in NN estimation (cf. Section 2.2.1). Particularly, $\sigma_{\mathcal{S}}(s_0, s_\imath) = 0$ means that the $\imath$-th case is not considered as a relevant piece of information since it is not sufficiently similar to $s_0$. For computation, irrelevant cases in (5.8) can clearly be left out of account. Thus, it is enough to consider cases in a certain region around $s_0$. As opposed to the $k$NN approach, it is the size of this region rather than the number of neighboring cases which is fixed.

As in previous chapters, we assume $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ to be reflexive and symmetric, whereas no special kind of transitivity is required.[10] In fact, the application of the maximum operator in (5.8) does even permit a purely *ordinal* approach. In this case, the range of the similarity measures is a finite subset $\mathcal{A} \subset [0, 1]$ that encodes an ordinal scale such as

$$\{\text{completely different}, \ldots, \text{very similar}, \text{identical}\}. \tag{5.10}$$

Correspondingly, degrees of possibility are interpreted in a qualitative way [251, 127]. That is, $\delta_{s_0}(r) < \delta_{s_0}(r')$ only means that outcome $r$ is less supported than outcome $r'$; apart from that, the difference between the possibility degrees has no meaning.

Needless to say, a scale such as (5.10) is more convenient if cases are complex objects rather than points in a Euclidean space and if similarity (distance) between objects must be assessed by human experts (which is common practice in case-based reasoning). Note that an ordinal structure is also sufficient for the original $k$NN rule. In connection with distance-weighting (cf. Section 2.2.1), however, the structures of the involved measures become more important. In any case, one should be aware of the fact that a cardinal interpretation of similarity raises some crucial semantic questions if corresponding measures cannot be defined in a straightforward way. In the weighted $k$NN rule, for example, one patient that died from a certain medical treatment compensates for two patients that survived if the former is twice as similar to the current patient. But what exactly does “twice as similar” mean in this context?

Looking at (5.8) from the point of view of observed cases, this estimation principle defines a (possibilistic) *extrapolation* of each case $\langle s_\imath, r_\imath \rangle$. In the original NN approach, which does not involve a distance measure on $\mathcal{R}$, a case $\langle s_\imath, r_\imath \rangle \in \mathcal{M}$ can only support the output $r_\imath$. This corresponds to the special case where $\sigma_{\mathcal{R}}$ in (5.8) is given by

[10] Let us mention again that relations satisfying reflexivity and symmetry are often called *proximity relations* in the fuzzy set literature, where similarity relations are defined as transitive proximity relations [100]. Anyway, we shall use the term similarity relation (similarity measure) henceforth without assuming transitivity.

$$\sigma_{\mathcal{R}}(r, r') = \begin{cases} 1 & \text{if} \quad r = r' \\ 0 & \text{if} \quad r \neq r' \end{cases}, \tag{5.11}$$

which is reasonable if $\mathcal{R}$ is a nominal scale, as, e.g., in concept learning.

By allowing for graded distances between outcomes, the possibilistic approach provides for a case $\langle s_\imath, r_\imath \rangle$ to support similar outcomes as well. This type of extended extrapolation is reasonable if $\mathcal{R}$ is a cardinal or at least ordinal scale. In fact, it should be observed that (5.8) applies to continuous scales in the same way as to discrete scales and thus unifies the performance tasks of classification and function approximation. For example, knowing that the price (= ouput) of a certain car is \$10,500, it is quite plausible that a similar car has exactly the same price, but it is plausible as well that it costs \$10,700. Interestingly enough, the same principle is employed in kernel-based estimation of probability density functions, where probabilistic support is allocated by kernel functions centered around observations [318, 289]. Indeed, (5.8) can be considered as a possibilistic counterpart of kernel-based density estimation. Let us furthermore mention that the consideration of graded distances between outputs is also related to the idea of class-dependent misclassification costs [290, 364].

## 5.3 Generalized possibilistic prediction

The possibility distribution $\delta_{s_0}$, which specifies the fuzzy set of well-supported outputs, is a disjunctive combination of the individual support functions

$$\delta_{s_0}^{\imath} : r \mapsto \min\left\{\sigma_{\mathcal{S}}(s_0, s_\imath),\, \sigma_{\mathcal{R}}(r, r_\imath)\right\}. \tag{5.12}$$

In fact, the max-operator in (5.8) is special t(riangular)-conorm and serves as a generalized logical or-operator: $r_0 = r$ is regarded as possible if $\langle s_0, r \rangle$ is similar to $\langle s_1, r_1 \rangle$ OR to $\langle s_2, r_2 \rangle$ OR ... OR to $\langle s_n, r_n \rangle$.

Now, fuzzy set theory offers t-conorms other than max and, hence, (5.8) could be generalized as follows:

$$\begin{aligned} \delta_{s_0}(r) &\overset{\text{df}}{=} \delta_{s_0}^{1}(r) \oplus \delta_{s_0}^{2}(r) \oplus \ldots \oplus \delta_{s_0}^{n}(r) \\ &= \bigoplus_{1 \leq \imath \leq n} \min\left\{\sigma_{\mathcal{S}}(s_0, s_\imath),\, \sigma_{\mathcal{R}}(r, r_\imath)\right\} \\ &= 1 - \bigotimes_{1 \leq \imath \leq n} \max\left\{1 - \sigma_{\mathcal{S}}(s_0, s_\imath),\, 1 - \sigma_{\mathcal{R}}(r, r_\imath)\right\} \end{aligned}$$

for all $r \in \mathcal{R}$, where $\otimes$ and $\oplus$ are a t-norm and a related t-conorm, respectively. Recall that a t-norm is a binary operator $\otimes : [0,1]^2 \longrightarrow [0,1]$ which is commutative, associative, monotone increasing in both arguments and which satisfies the boundary conditions $x \otimes 0 = 0$ and $x \otimes 1 = x$ [227]. An associated t-conorm is defined by the mapping $(\alpha, \beta) \mapsto 1 - (1-\alpha) \otimes (1-\beta)$. The t-norm associated with the t-conorm max is the min-operator. Other important operators are the product $\otimes_P : (\alpha, \beta) \mapsto \alpha\beta$ with related t-conorm $\oplus_P : (\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ and the Lukasiewicz t-norm $\otimes_L : (\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$ the related t-conorm of which is the bounded sum $\oplus_L : (\alpha, \beta) \mapsto \min\{1, \alpha + \beta\}$.

Observe that the minimum operator employed in the determination of the joint similarity between cases can be considered as a logical operator as well, namely as a fuzzy conjunction: Two cases $\langle s_0, r \rangle$ and $\langle s_\imath, r_\imath \rangle$ are similar if both, $s_0$ is similar to $s_\imath$ *and* $r$ is similar to $r_\imath$. Consequently, this operator might be replaced by a t-norm, too. By doing so, (5.12) and (5.8) become

$$\delta_{s_0}^{\imath} : r \mapsto \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath) \tag{5.13}$$

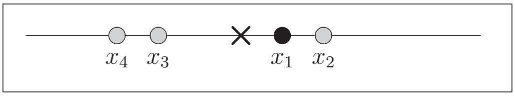
and

$$\delta_{s_0}(r) \overset{\text{df}}{=} \bigoplus_{1 \leq \imath \leq n} \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath), \tag{5.14}$$

respectively. Note, however, that a (fuzzy) logic-based derivation of the joint similarity is not compulsory. Particularly, the t-norm $\otimes$ in (5.14) need not necessarily be the one related to the t-conorm $\oplus$. For example, one might thoroughly take $\otimes = \min$ and $\oplus = \oplus_P$, or even combine the similarity degrees $\sigma_{\mathcal{S}}(s_0, s_\imath)$ and $\sigma_{\mathcal{R}}(r, r_\imath)$ by means of an operator which is not a t-norm. In that case, however, the "logical" interpretation of (5.14) is lost.

### 5.3.1 Control of compensation and accumulation of support

By choosing an appropriate t-conorm $\oplus$ in (5.14) one can control the accumulation of individual degrees of evidential support, especially the extent of compensation. To illustrate, consider the following classification scenario (with labels DARK and LIGHT), where $\sigma_{\mathcal{S}}(s_0, s_1) = 3/4$, $\sigma_{\mathcal{S}}(s_0, s_2) = \sigma_{\mathcal{S}}(s_0, s_3) = 1/2$, and $\sigma_{\mathcal{S}}(s_0, s_4) = 1/4$:



Should one prefer DARK or LIGHT as a classification of the new input (indicated by the cross)? The use of the max-operator as a t-conorm yields $\delta_{s_0}(\text{DARK}) = 3/4$ and $\delta_{s_0}(\text{LIGHT}) = 1/2$ and, hence, the decision DARK. The three moderately similar instances with label LIGHT do not compensate for the one very similar instance with label DARK. As opposed to this, the probabilistic sum $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ brings about a compensation effect and entails $\delta_{s_0}(\text{DARK}) = 3/4$ and $\delta_{s_0}(\text{LIGHT}) = 13/16$, that is, a slightly larger possibility for LIGHT.

More generally, different t-conorms can model different accumulation modes, which typically entail a kind of saturation effect. In the case of the probabilistic sum $\oplus_P$, for example, an additional $\beta$-similar observation increases the current support $\alpha$ by $\beta(1-\alpha)$. Thus, the larger the support already granted is, the smaller the absolute increase due to the new observation will be. This appears reasonable from an intuitive point of view: If the support of an output is already large, one is not surprised to see another (close) input having the same output. A small support increment then reflects the low information content related to the new observation [203].

### 5.3.2 Possibilistic support and weighted NN estimation

A t-norm $\otimes$ is called Archimedian if the following holds: For all $x, y \in ]0,1[$ there is a number $n \in \mathfrak{N}$ such that $\otimes^{(n)}(x) < y$ (where $\otimes^{(n)}(x) = \otimes^{(n-1)}(x) \otimes x$ and $\otimes^{(1)}(x) = x$). It can be shown that $\otimes$ is a continuous Archimedian t-norm iff there is a continuous, strictly decreasing function $g : [0,1] \longrightarrow [0,\infty]$ such that $g(1) = 0$ and

$$\alpha \otimes \beta = g^{(-1)}(g(\alpha) + g(\beta)) \tag{5.15}$$

for all $0 \leq \alpha, \beta \leq 1$, where the pseudo-inverse $g^{(-1)}$ is defined as

$$g^{(-1)} : x \mapsto \begin{cases} g^{-1}(x) & \text{if} \quad 0 \leq x \leq g(0) \\ 0 & \text{if} \quad g(0) < x \end{cases} .$$

The function $g$ is called the *additive generator* of $\otimes$. For example, $x \mapsto 1 - x$ and $x \mapsto -\ln(x)$ are additive generators of the Lukasiewicz t-norm $\otimes_L$ and the product $\otimes_P$, respectively.

Based on the representation (5.15), one can establish an interesting connection between (5.14) and the weighted NN rule (cf. Section 2.2.1). To this end, let $g$ be the additive generator of the t-norm[11] related to the t-conorm $\oplus$ used as an aggregation operator in (5.14). With $d_\imath = 1 - \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath)$ and $\omega_\imath = g(d_\imath)$, we can write (5.14) as

$$\delta_{s_0}(r) = 1 - g^{(-1)}(\omega_1 + \omega_2 + \ldots + \omega_n). \tag{5.16}$$

Since $g$ is decreasing, it can be considered as a weight function that turns a distance $d_\imath$ into a weight $\omega_\imath$ associated with the $\imath$-th input. Then, (5.16) tells us that the possibility degree $\delta_{s_0}(r)$ is nothing else than a (monotone increasing) transformation of the sum of weights $\omega_\imath$. In other words, (5.14) can be seen as a distance-weighted NN estimation, where the weight of a neighbor is determined as a function of its similarity to the new instance. As opposed to (2.8), however, the weight of a case according to (5.16) does not depend on other cases stored in the memory (cf. Section 5.3.5 below).

Consider the Lukasiewicz t-(co)norm as an example, for which we obtain $\omega_\imath = 1 - d_\imath = \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath)$ and

[11] This is not the t-norm used in (5.14) for defining a joint similarity measure.

$$\delta_{s_0}(r) = \min\{1, \omega_1 + \omega_2 + \ldots + \omega_n\}. \tag{5.17}$$

If, moreover, $\sigma_{\mathcal{R}}$ is given by (5.11), then $\delta_{s_0}(r)$ is nothing else than the bounded sum of the similarity degrees $\sigma_{\mathcal{S}}(s_\imath, s_0)$ between $s_0$ and the inputs $s_\imath$ with output $r_\imath = r$. Thus, (5.17) is basically equivalent to the global NN method, i.e., the weighted NN approach with $k = n$,[12] apart from the fact that it does not distinguish between outputs whose accumulated support exceeds 1 (this is an extreme type of saturation effect). For the probabilistic sum $\oplus_P$, the mapping between possibility degrees and the sum of weights is bijective:

$$\delta_{s_0}(r) = 1 - \exp\big(-(\omega_1 + \omega_2 + \ldots + \omega_n)\big).$$

In connection with the generalized model (5.14), the t-conorm $\oplus$ used for combining individual degrees of support defines another degree of freedom of the model. It is hence interesting to mention the existence of parameterized families of t-(co)norms which comprise commonly used operators as special cases. For example, the Frank-family is defined as

$$\oplus_\rho : (\alpha, \beta) \mapsto \begin{cases} \max\{\alpha, \beta\} & \text{if } \rho = 0 \\ \alpha + \beta - \alpha\beta & \text{if } \rho = 1 \\ \min\{1, \alpha + \beta\} & \text{if } \rho = \infty \\ 1 - \ln_\rho\left(1 + \frac{(\rho^{1-\alpha}-1)(\rho^{1-\beta}-1)}{\rho - 1}\right) & \text{otherwise} \end{cases} . \tag{5.18}$$

Proceeding from such a family of t-conorms, the degree of freedom of the model reduces to a single parameter, here $\rho$, which can be adapted in a simple way, e.g., by means of cross-validation techniques.

### 5.3.3 Upper and lower possibility bounds

The possibility degree (5.14) represents the support (confirmation) of an output $r$ gathered from similar cases according to the CBI hypothesis. Now, in the sense of this hypothesis, an observation $\langle s_\imath, r_\imath \rangle$ might not only confirm but also *disqualify* an output $r$. This happens if $s_\imath$ is close to $s_0$ but $r_\imath$ is not similar to $r$. A possibility distribution expressing degrees of *exclusion* rather than degrees of support and, hence, complementing (5.14) in a natural way is given by

$$\pi_{s_0} : r \mapsto \bigotimes_{1 \le \imath \le n} (1 - \sigma_{\mathcal{S}}(s_0, s_\imath)) \oplus \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.19}$$

According to (5.19), an individual observation $\langle s_\imath, r_\imath \rangle$ induces a constraint on the outcome of $s_0$: An output $r$ is disqualified by $\langle s_\imath, r_\imath \rangle$ if both, $\sigma_{\mathcal{S}}(s_0, s_\imath)$ is large and $\sigma_{\mathcal{R}}(r, r_\imath)$ is small. As opposed to this, $\langle s_\imath, r_\imath \rangle$ is completely ignored if

[12] The proper $k$NN rule cannot be emulated as in (2.10) since the weights $\omega_\imath$ depend on absolute distance (again, see Section 5.3.5 below).

$\sigma_{\mathcal{S}}(s_0, s_\imath) = 0$, in which case the individual support on the right-hand side of (5.19) is 1 ($\pi_{s_0} \equiv 1$ is an expression of complete ignorance: all upper possibility bounds are 1 since there is no reason to discredit any output). This approach is obviously in agreement with the constraint-based view of possibilistic reasoning (cf. Section 5.1.1). Moreover, the distribution (5.19) is again related to a special type of fuzzy rule [107].

The possibility of an outcome $r$ can now be characterized by means of an extended estimation, namely as a tuple

$$\delta^*_{s_0}(r) = [\, \delta_{s_0}(r),\, \pi_{s_0}(r)\, ]$$

with a lower bound $\delta_{s_0}(r)$ expressing a degree of confirmation, and an upper bound $\pi_{s_0}(r)$ expressing a degree of plausibility. The following cases show that the complementary distribution $\pi_{s_0}$ can greatly improve the informational content of a possibilistic evaluation:[13]

- $\delta^*_{s_0}(r) = [0, 1]$: This is an expression of complete ignorance. Neither is $r$ supported nor is it (partly) excluded by any observation. Thus, $r$ is fully plausible though not confirmed at all.
- $\delta^*_{s_0}(r) = [0, 0]$: Clear evidence against $r$ has been accumulated in the form of inputs similar to $s_0$ with outputs dissimilar to $r$.
- $\delta^*_{s_0}(r) \approx [1, 1]$: The output $r$ is strongly supported through the observation of similar cases.

Notice that

$$\delta_{s_0}(r) > \pi_{s_0}(r) \tag{5.20}$$

indicates a kind of conflict [376] and is closely related to the problem of ambiguity in connection with the NN principle (cf. Section 2.2.1). In fact, (5.20) can occur if $s_0$ has close neighbors $s_\imath$ and $s_\jmath$ with quite dissimilar outputs $r_\imath$ and $r_\jmath$ (mathematically speaking, $s_0$ is a point of discontinuity). In this case, the evaluation of $r$ is unsteady, and the support $\delta_{s_0}(r)$ should be taken with caution. The inequality in (5.20) might also trigger a revision process that aims at removing the conflict by means of a model adaptation.

### 5.3.4 Fuzzy logical evaluation

The values $\delta_{s_0}(r)$ in (5.14) can also be considered as membership degrees of a fuzzy set, namely the fuzzy set of "well-supported outputs". In fact, the possibility degree $\delta_{s_0}(r)$ can be seen as the truth degree, $\langle P(r) \rangle$, of the following (fuzzy) predicate $P(r)$: "There is an input close to $s_0$ with an output similar to $r$." $P(r)$ defines the property that qualifies $r$ as a well-supported output.

[13] Recall that positive and negative evidence cannot be distinguished in probability theory.

Of course, one might easily think of alternative characterizations of well-supported outputs. Fuzzy set-based modeling techniques allow for translating such characterizations given in linguistic form into logical expressions. By using fuzzy logical connectives including t-norms, fuzzy quantifiers such as "a few" and fuzzy relations such as "closely located", one can specify sophisticated fuzzy decision principles that go beyond the simple NN rule. Example:

> "There are at least a few closely located inputs, most of these inputs have the same output, and none of the moderately close inputs has a very different output."

The logical expression $P(\cdot)$ associated with such a specification can be used in place of the right-hand side in (5.14):

$$\delta_{s_0}(r) \overset{\text{df}}{=} \langle P(r) \rangle. \tag{5.21}$$

The decision rule related to (5.14) favors the outcome $r_0^{est}$ that meets the requirements specified by $P(\cdot)$ best. This generalization appears especially interesting since it allows one to adapt the NN principle so as to take specific characteristics of the application into account.

Observe that (5.21) can also mimic the original $k$NN rule: Consider the fuzzy proposition "$r$ is supported by many of the $k$ nearest neighbors of $s_0$", and let the fuzzy quantifier "many (out of $k$)" be modeled by the mapping $\imath \mapsto \imath/k$. Then, $\delta_{s_0}(r) = \imath/k$ iff $\imath$ among the $k$ nearest neighbors have outcome $r$. In this case, possibility degrees (derived from fuzzy truth degrees) formally coincide with probability degrees.

### 5.3.5 Comparison of extrapolation principles

As already mentioned above, the possibilistic approach to CBI can also be considered as a kind of NN estimation. Thus, it seems interesting to have a closer look at this type of "possibilistic NN estimation" as an alternative to the *probabilistic* approach to estimation and decision making, which is in agreement with the original $k$NN rule (cf. Section 2.2.1).

Both the possibilistic and the probabilistic approach can be considered as a two-step procedure. The first step derives a distribution that will subsequently be referred to as the NN *estimation*. This estimation defines a degree of support for each output $r \in \mathcal{R}$. The second step, the NN *decision*, chooses one output on the basis of the NN estimation. Usually, the decision is given by the outcome with maximal support, and ties are broken by coin flipping. Still, in the case of a continuous (or at least ordinal) scale $\mathcal{R}$, a decision might also be obtained by some kind of averaging procedure.

In order to facilitate the comparison of the two approaches, we write degrees of evidential support in the general form

$$\nu(r \,|\, s_0, \mathcal{M}) = \alpha\left(\{\nu_{s_\imath}(r \,|\, s_0, \mathcal{M}) \,|\, \langle s_\imath, r_\imath \rangle \in \mathcal{M}\}\right) \tag{5.22}$$

and thus obtain the (maximal support) decision as

$$r_0^{est} = \arg\max_{r \in \mathcal{R}} \nu(r \,|\, s_0, \mathcal{M}). \tag{5.23}$$

In (5.22), $\nu_{s_\imath}(r \,|\, s_0, \mathcal{M})$ is the support of the hypothesis $r_0 = r$ provided by the case $\langle s_\imath, r_\imath \rangle$, and $\alpha$ is an aggregation function.

To reveal the original $k$NN rule and the probabilistic approach as special cases of (5.23), note that the probability distribution (2.6) is obtained by using the arithmetic sum as an aggregation function $\alpha$ and defining the support function as

$$\nu^p_{s_\imath}(r \,|\, s_0, \mathcal{M}) = \left\{ \begin{array}{cl} 1/k & \text{if } s_\imath \in \mathcal{N}_k(s_0) \text{ and } r = r_\imath \\ 0 & \text{otherwise} \end{array} \right. . \tag{5.24}$$

More generally, if $\mathcal{S}$ is a metric space, a support function can be defined as

$$\nu^p_{s_\imath}(r \,|\, s_0, \mathcal{M}) = \left\{ \begin{array}{cl} K_{d_k}(s_0 - s_\imath) & \text{if } r = r_\imath \\ 0 & \text{otherwise} \end{array} \right. , \tag{5.25}$$

where $K$ is a kernel function. The index $d_k$ denotes the distance between $s_0$ and its $k$-th nearest neighbor. It signifies that the kernel function is *scaled* so as to exclude exactly those inputs $s_\imath$ with $\Delta_{\mathcal{S}}(s_0, s_\imath) > d_k$. Proceeding from (5.25), and assuming that $\mathcal{R}$ is a finite set $\{\rho_1 \ldots \rho_m\}$, the probability distribution $p_{s_0}$ is obtained by normalizing the supports

$$\nu^p(\rho_\jmath \,|\, s_0, \mathcal{M}) = \sum_{\langle s_\imath, r_\imath \rangle \in \mathcal{M}} \nu^p_{s_\imath}(\rho_\jmath \,|\, s_0, \mathcal{M}),$$

which yields

$$p_{s_0}(\rho_\jmath) = \frac{\nu^p(\rho_\jmath \,|\, s_0, \mathcal{M})}{\sum_{\imath=1}^m \nu^p(\rho_\imath \,|\, s_0, \mathcal{M})} \tag{5.26}$$

for all $\rho_\jmath \in \mathcal{R}$. That is, the aggregation $\alpha$ is now the normalized rather than the simple arithmetic sum. Of course, since normalization does not change the mode of a distribution it has no effect on decision making and could hence be omitted from this point of view.

The possibilistic approach (5.14) is recovered by $\alpha = \oplus$ and

$$\nu^\delta_{s_\imath}(r \,|\, s_0, \mathcal{M}) = \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.27}$$

As can be seen, the main difference between the probabilistic and the possibilistic approach concerns the definition of the individual support function $\nu_s$ and the aggregation of the corresponding degrees of support.

Apart from that, however, a direct comparison is complicated by the similarity measure over outputs, $\sigma_{\mathcal{R}}$, which is used in (5.27) but not in (5.25). One possibility to handle this problem is to consider (5.27) only for the special case (5.11):

$$\nu^{\delta}_{s_\imath}(r \,|\, s_0, \mathcal{M}) = \begin{cases} \sigma_{\mathcal{S}}(s_0, s_\imath) & \text{if } r = r_\imath \\ 0 & \text{otherwise} \end{cases} . \tag{5.28}$$

Equation (5.28) reveals that the similarity measure $\sigma_{\mathcal{S}}$ now plays the same role as the kernel function $K$ in (5.25).

**Absolute versus relative support.** An important difference between (5.25) and (5.28) is that an example $\langle s_\imath, r_\imath \rangle \in \mathcal{M}$ provides *relative* support of an output $r$ in the probabilistic approach but *absolute* support in the possibilistic one. That is, $\nu^{\delta}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ depends on the absolute similarity between $s_0$ and $s_\imath$ but is independent of further observations. In fact, we can actually write $\nu^{\delta}_{s_\imath}(r \,|\, s_0)$ in place of $\nu^{\delta}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ since $\mathcal{M}$ does not appear on the right-hand side of (5.28): The support provided by observed examples $\langle s_\imath, r_\imath \rangle$ is bounded to nearby cases, decreases gradually with distance, and vanishes for completely dissimilar cases.

As opposed to this, the support $\nu^{p}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ is relative and depends on the relation between the distance of $s_\imath$ to $s_0$ and the distances of other observations to $s_0$. This is reflected by the scaling of the kernel function in (5.25). On the one hand, this means that $\nu^{p}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ can be large even though $s_\imath$ is quite distant from $s_0$. On the other hand, the extension of the memory $\mathcal{M}$ by another instance close enough to $s_0$ might exclude a quite similar observation $s_\imath$ from the neighborhood $\mathcal{N}_k(s_0)$. The corresponding re-scaling of the kernel function will then cancel the support provided by $\langle s_\imath, r_\imath \rangle$ so far. The induced thresholding effect appears especially radical (and might be questioned on such grounds) in connection with (5.24), where $\nu^{p}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ is reduced from $1/k$ to 0, that is from full support to zero support.

The bounding of evidential support, as realized by the possibilistic approach, is often advisable. Consider a simple example: Let $\mathcal{S} = [0, 1]$ and

$$\varphi = \{(s, \mathbb{I}_{[1/2,1]}(s)) \,|\, s \in \mathcal{S}\}$$

and suppose inputs to be chosen at random according to a uniform distribution. Moreover, assume that a new input $s_0$ must be labeled, given a memory that consists of only a single observation $\langle s_1, r_1 \rangle$. Using the 1NN rule, the probability of a correct decision is obviously 1/2. Now, suppose that the NN rule is applied only if $|s_0 - s_1| \le d$, whereas a decision is determined by flipping a coin otherwise (this is exactly the procedure that results from the possibilistic approach by defining $\sigma_{\mathcal{S}}$ in (5.8) by $\sigma_{\mathcal{S}}(s, s') = 1$ if $|s - s'| \le d$ and 0 otherwise). A simple calculation shows that the probability of a correct decision is now $1/2 + d(1 - d)$. As can be seen, dissimilar cases are likely to provide misleading information in this example and, hence, the disregard of such cases is indeed advantageous. Loosely speaking, it is better to guess an output at random than to rely on observations not similar enough.

Of course, the concept of absolute support is actually not reserved to the possibilistic approach but can be realized for the probabilistic method as well. To this end, one simply replaces (5.25) by

$$\nu^p_{s_\imath}(r \,|\, s_0, \mathcal{M}) = \begin{cases} K(s_0 - s_\imath) & \text{if } r = r_\imath \\ 0 & \text{otherwise} \end{cases} , \tag{5.29}$$

where the kernel function $K$ is now fixed. That is, $K$ is no longer scaled by the size of the neighborhood of $s_0$. This is exactly the estimation one derives by the reasoning in Section 2.2.1 if the generalized NN density estimation (2.14) is replaced by the simple kernel estimator:

$$\phi^{est}(s_0) = \frac{1}{n} \cdot \sum_{\imath=1}^{n} K(s_0 - s_\imath). \tag{5.30}$$

Here, the only problem occurs if $\nu^p(r \,|\, s_0, \mathcal{M}) = 0$ for all $r \in \mathcal{R}$. In this situation (of complete ignorance), a probability distribution cannot be derived by normalization.

Apart from that, (5.29) might indeed be preferred to (5.25) due to the reasons mentioned above. In fact, one should realize that one of the major reasons for using the NN density estimator (2.14) rather than the kernel estimator (5.30) is to guarantee the continuity of the density function $\phi^{est}$. In the context of case-based inference or, say, instance-based learning this is not important, however, since one is not interested in estimating a complete density function but only a single value thereof. To the best of our knowledge, (5.25) and (5.29) have not been compared in a systematic way in IBL so far. Note that (5.29) should actually be called a Near Neighbor estimation since it involves the *near* rather than the *nearest* neighbors. The same remark applies to the possibilistic approach, of course.

Above, it has been argued that the consideration of graded degrees of similarity between outcomes is often advised (see also our example in Section 5.3.7 below). It should be mentioned, therefore, that the probabilistic approach might be extended in this direction as well. To this end, a *joint* probability density can be estimated based on a kernel function $K$, which is now defined over $\mathcal{S} \times \mathcal{R}$. An estimation for the output $r$ can then be derived by conditioning on $s_0$:

$$p_{s_0}(r) \;\propto \sum_{\langle s_\imath, r_\imath \rangle \in \mathcal{M}} \nu^p_{s_\imath}(r \,|\, s_0, \mathcal{M}) \;=\; \sum_{\langle s_\imath, r_\imath \rangle \in \mathcal{M}} K\left(s_0 - s_\imath,\, r - r_\imath\right).$$

This is the most general form of a probabilistic estimation. Still, one should keep in mind that it requires $\mathcal{S} \times \mathcal{R}$ to have a suitable mathematical structure, an assumption which is not always satisfied in applications (again, we refer to our example below).

**Similarity versus frequency.** The estimation principle underlying the probabilistic NN approach combines the concepts of similarity (distance) and frequency: It applies a closeness assumption, typical of similarity-based reasoning, that suggests to focus on the most similar observations (or to weight observations by their distance). From the reduced set of supposedly most relevant instances, probabilities are then estimated by relative frequencies. This contrasts with the basic (max–min) possibilistic approach (5.8) which relies on similarity alone: The application of the maximum operator does not produce any compensation or reinforcement effect. Thus, possibility depicts the *existence* of supporting evidence, not its frequency.[14] The generalized possibilistic approach based on (5.14) allows for modes of compensation which combine both aspects. Especially, the operators mentioned above produce a kind of saturation effect, that is, a limited reinforcement effect: The increase of support due to the observation of a similar instance is a decreasing function of the support that is already available.

In this connection, it is important to realize the different nature of the concepts of possibility and probability. Particularly, it should be emphasized that the former is not interpreted in terms of the latter.[15] For example, consider the standard probabilistic setting where cases are chosen randomly and independently according to a fixed probability measure over $\mathcal{S} \times \mathcal{R}$. The possibility degree $\delta_{s_0}(r)$ will then converge to 1 with increasing sample size whenever $\langle s_0, r \rangle$ has a non-zero probability of occurrence. In fact, the possibilistic approach is interested in the *existence* of a case, not in its probability. Roughly speaking, the major concern of this approach is the approximation of the concepts $C_r$, $r \in \mathcal{R}$, whereas the probabilistic approach aims at estimating conditional probability distributions $p_{s_0} = \mathbb{P}(\cdot \,|\, s_0)$. Of course, this distinction is relevant only if the concepts are overlapping, that is, if the query $s_0$ does not have a unique outcome. Otherwise, a possibilistic and a probabilistic approach are equivalent in the sense that $s_0 \in C_r \Leftrightarrow \mathbb{P}(r \,|\, s_0) = 1$.

It is beyond question that the frequency of observations usually provides valuable information. Yet, the frequency-based approach does heavily rely on statistical assumptions concerning the generation of training (and test) data. Thus, it might be misleading if these assumptions are violated. Suppose, e.g., that the probability of observing a positive example, while learning a concept $C_1 \subseteq \mathcal{S}$, depends on the number of positive examples observed so far and hence contradicts an independence assumption (the probability of an output $r$, given the input $s$, is not independent of the data). In this case, a probabilistic estimation is clearly biased, whereas the possibility distribution (5.8) is not affected at all. Indeed, the information expressed by $\delta_{s_0}$ remains valid even if only negative examples $s_\imath \in C_0 = \mathcal{S} \setminus C_1$ have been presented so far: $\delta_{s_0}(1) = 0$ then simply means that no evidence for $s_0 \in C_1$ has been gathered as yet. Moreover, the value $\delta_{s_0}(0)$ reflects the available support for $s_0 \in C_0$. This support depends on the distance of $s_0$ to the observed negative examples. Note that $\delta_{s_0}(0) = 0$ is possible as well. In this case, no evidence is available at all, neither for nor against $s_0 \in C_1$. See Section 5.5.3 for a simulation experiment which concerns the aspect of robustness of NN estimation toward violations of the standard statistical assumptions.

[14] To a certain extent, this is related to the distinction between an *existential* and an *enumerative* analogy factor in models of analogical induction [281].

[15] Though such a relationship can be established, e.g., by interpreting possibility as upper probability [122] or fuzzy sets as coherent random sets [111].

Apart from statistical assumptions, the structure of the application has an important influence. To illustrate, consider two classes in the form of two clusters such that the (known) diameter of both clusters is smaller than the distance between them, that is $\varDelta_{\mathcal{S}}(s_1, s_2) < \varDelta_{\mathcal{S}}(s_1, s_3)$ whenever $r_1 = r_2 \neq r_3$. The output of an input can then be determined with certainty as soon as the distance from its nearest neighbor is known. In other words, the 1NN rule which does not involve frequency information performs better than any $k$NN rule with $k > 1$.

### 5.3.6 From predictions to decisions

In addition to the extrapolation principles let us compare the induced distributions, referred to as NN estimations, from a knowledge representational point of view, especially against the background of the two shortcomings of the NN rule illustrated in Fig. 2.1.

A crucial difference between a possibility distribution $\delta$ and a probability function $p$ is that the latter obeys a normalization constraint that demands a total probability mass of 1, whereas no such constraint exists in possibility theory. Consequently, a possibility distribution is more expressive in some situations. Especially, the following points deserve mentioning:

- Possibility reflects ignorance: All possibility degrees $\delta_{s_0}(r)$ remain rather small if no sufficiently similar cases are available. Particularly, the distribution $\delta_{s_0} \equiv 0$ is an expression of *complete ignorance* and reflects the absence of any relevant observation ($\sigma_{\mathcal{S}}(s_0, s_\imath) = 0$ for all $s_\imath$). A learning agent using this estimation "knows that it doesn't know" [359]. As opposed to this, a distribution such as, say, $\delta_{s_0} \equiv |\mathcal{R}|^{-1}$ (in the case of finite $\mathcal{R}$) indicates that some (small) evidence is available for each of the potential outcomes. These two situations cannot be distinguished in probability theory where they induce the same distribution $p_{s_0} \equiv |\mathcal{R}|^{-1}$ (if, as suggested by the principle of insufficient reason, complete ignorance is modeled by the uniform distribution).
- Possibility reflects absolute frequency: For example, suppose $\sigma_{\mathcal{S}}(s_0, s_\imath) = 1 - d > 0$ and $r_\imath = r'$ for all $n$ inputs $s_\imath$ stored in the memory. The probabilistic estimation (2.6) then yields the one-point distribution $p_{s_0}(r') = 1$ and $p_{s_0}(r) = 0$ for all $r \neq r'$. Thus, it suggests that $r_0 = r'$ is certain, even if $n$ is rather small. With a compensating t-conorm such as the probabilistic sum $\oplus_P$, the extended estimation (5.14) yields $\delta_{s_0}(r') = 1 - d^n$ and $\delta_{s_0}(r) = 0$ for all $r \neq r'$. Thus, not only does the possibilistic support of the hypothesis $r_0 = r'$ reflect the distance but also the actual number of voting instances: $\delta_{s_0}(r')$ is an increasing function of $n$ and approaches 1 for $n \to \infty$.

As can be seen, a probabilistic estimation can represent ambiguity, whereas the possibilistic approach captures both problems, ambiguity and ignorance: Ambiguity (Fig. 2.1, above) is present if there are several plausible outputs with similar degrees of support, and ignorance (Fig. 2.1, below) is reflected by the fact that even the most supported output has a small degree of possibility. Thus, (5.14) can be taken as a point of departure for a decision making procedure that goes beyond the guessing of an outcome. For example, a possible line of action proceeding from (5.14) might be expressed by the following rules (involving thresholds $0 < d_{max} < d_{min} < 1$):

- If $\delta_{s_0}(r^*) \geq d_{min}$ for the most supported outcome $r^*$ and $\delta_{s_0}(r) \leq d_{max}$ for all $r \neq r^*$, then let $r_0^{est} = r^*$.
- If $\delta_{s_0}(r^*) < d_{min}$, then gather further information.
- If $\delta_{s_0}(r^*) \geq \delta_{s_0}(r) \geq d_{min}$ for two outcomes $r^*, r \in \mathcal{R}$, then refuse a prediction.

The ECHOCARDIOGRAM DATABASE[16] is a real-world example that is quite interesting in this respect. One problem that has been addressed by machine learning researchers in connection with this database is to predict from several attributes whether or not a patient who suffered from a heart attack will survive at least one year. Since data is rather sparse (132 instances and about 10 attributes), the possibilistic approach often yields estimations with low support for both alternatives, surviving and not surviving at least one year. This is clearly reasonable from a knowledge representational point of view and reveals an advantage of absolute over relative degrees of support. For example, telling a patient that your experience does not allow any statement concerning his prospect of survival ($\delta_{s_0} \equiv 0$) is very different from telling him that his chance is 50% ($p_{s_0} \equiv 1/2$).

The discrepancy between a probabilistic and a possibilistic approach disappears to some extent if one is only interested in a final decision, that is, if a decision must be made irrespective of the quality and quantity of the information at hand. For example, the method in [84], which derives a prediction in terms of a *belief function* (cf. Chapter 4), refers to the so-called transferable belief model [350] and, hence, turns the belief function (at the "credal" level) specifying the unknown outcome into a probability function (at the "pignistic" level) before making a decision. Thus, the support of individual outputs is expressed in terms of probability, and an NN estimation can be derived by taking one among the most probable outcomes, breaking ties at random.

Observe that, as a consequence of applying the maximum operator, a possibilistic NN decision derived from (5.8) coincides with the 1NN rule. The generalized version (5.14), where several moderately similar examples can compensate for one very similar instance, comes closer to the original $k$NN rule. In fact, for certain special cases, the possibilistic approach is equivalent – from a decision making point of view – to the probabilistic approach based on the support function (5.29). Equation (5.16) shows that a possibility degree $\delta_{s_0}(r)$ is a monotone transformation of the sum of weights $\omega_\imath$, and this relation is one-to-one if the pseudo-inverse $g^{(-1)}$ is actually the inverse $g^{-1}$. The similarity function $\sigma_{\mathcal{S}}$ can then be chosen

[16] Available at http://www.ics.uci.edu/~mlearn.

such that

$$\delta_{s_0}(r) \leq \delta_{s_0}(r') \;\Leftrightarrow\; p_{s_0}(r) \leq p_{s_0}(r').$$

That is, outcomes which are better supported in a possibilistic sense are also more probable and vice versa.

To illustrate, consider the case where $\mathcal{S} = \mathfrak{R}^l$ and $\sigma_{\mathcal{R}}(r, r') = 1$ if $r = r'$ and 0 otherwise. Let $K$ be a kernel function and define $\sigma_{\mathcal{S}}$ as $(x, y) \mapsto 1 - \exp(-K(x, y))$.[17] For the t-conorm $\oplus_P$, the weights in (5.16) are then given by $\omega_\imath = K(s_0 - s_\imath)$. Therefore,

$$\begin{aligned}\delta_{s_0}(r) &= 1 - \exp\left( - \sum_{\langle s_\imath, r_\imath\rangle \in \mathcal{M} \,:\, r_\imath = r} K(s_0 - s_\imath) \right) \\ &= 1 - \exp\left(-c \cdot p_{s_0}(s_\imath)\right),\end{aligned}$$

where $p_{s_0}(r)$ is the probability degree derived from (5.29) using the kernel function $K$ and $c$ is the normalization factor $c = \sum_{r' \in \mathcal{R}} p_{s_0}(r')$.

### 5.3.7 An illustrative example

Here, we present a simple example for which the possibilistic approach might be considered superior to the probabilistic one. The task shall be to predict a student's grade in physics given some information on other grades of that student. Thus, an input is now a subject, and the output is given by the corresponding grade. We assume that grades are taken from the scale $\mathcal{R} = \{0, 1, \ldots, 10\}$, where 10 is the best result. Moreover, we consider two scenarios S1 and S2:

| Subject | S1 | S2 |
|---|---|---|
| Chemistry | – | 10 |
| French | – | 3 |
| Philosophy | – | 3 |
| Spanish | – | 3 |
| Sports | 5 | – |

Admittedly, it is not obvious how to define a reasonable similarity measure over the set of subjects. In fact, an ordinal measure – sufficient for the possibilistic approach (5.8) – appears much simpler than a cardinal one. Nevertheless, let us assume the following (cardinal) degrees of similarity:

| $\sigma_{\mathcal{S}}$ | Chem. | French | Phil. | Span. | Sports |
|---|---|---|---|---|---|
| Physics | 3/4 | 1/3 | 1/3 | 1/3 | 0 |

[17] Formally, one might set $K(0) \overset{\text{df}}{=} \infty$ to ensure that $\sigma_{\mathcal{S}}$ is reflexive.

Concerning the set of outcomes $\mathcal{R}$, graded degrees of similarity are clearly advised in this example. Let us define the similarity between two grades $a$ and $b$ to be

$$\sigma_{\mathcal{R}}(a,b) = \max\left\{1 - \frac{1}{5}|a-b|, 0\right\}.$$

Needless to say, our application does not define a statistical setup par excellence, which is a main reason why the probabilistic approach does hardly appear suitable. To begin with, a scenario as defined above cannot be considered as an independent sample (perhaps the information is censored if it comes from the student himself), not to mention the small number of observations. Moreover, a relative frequency interpretation does not make sense. Finally, the set $\mathcal{S}$ endowed with the similarity measure $\sigma_{\mathcal{S}}$ (as partly specified above) is likely to lack a sufficiently strong mathematical (metric) structure, so that the derivation of the $k$NN estimation in Section 2.2.1 might no longer be valid. Clearly, nothing prevents us from still applying the formulae and simply interpreting the normalized degrees of additive support as degrees of probability. But one should keep in mind that this approach actually lacks a solid foundation.

The first scenario is a typical example of complete ignorance, for one does not have any relevant piece of information. It is true that the case base is not empty, but the grade in sports does not allow one to draw any conclusion on the grade in physics since these two subjects are very dissimilar. This is adequately reflected by the possibilistic estimation which yields $\delta_{s_0} = \delta_{physics} \equiv 0$. A probabilistic estimation with relative support is obviously not appropriate in this example. Since sports is the only neighbor one obtains a probability distribution that favors grade 5 for physics. Thus, it is clearly advised to use absolute rather than relative support. Then, however, a probability is actually not defined since the denominator in (5.26) is zero. One way out is to take the uniform distribution $p_{s_0} \equiv 1/11$ as a default estimation, but this raises the well-known question whether the latter is an adequate expression of complete ignorance (which is definitely denied by most scholars).

Scenario S2 reveals problems of weighting and aggregation. Undoubtedly, a weighted estimation should be preferred in this example. Still, the example shows that the definition and aggregation of weights can be tricky. What is the most likely grade? Particularly, is grade 3 for physics more likely than grade 10 or vice versa? The weighted $k$NN rule favors grade 3 since the three subjects which are moderately similar to physics compensate for the one (chemistry) which is very similar. Of course, this result might be judged critically. Especially, this example reveals a problem of interdependence which is not taken into account by means of a simple summation of weights. Namely, the two subjects Spanish and French are very similar by themselves. Thus, one might wonder whether the grade 3 should really count twice. In fact, one might prefer to consider the grades in French and Spanish as only one piece of evidence (suggesting that the student is not good at languages) instead of two pieces of distinct information. Formally, the problem is that the probabilistic approach makes an assumption of (conditional) independence which is no longer valid when taking *structural* assumptions about the application into account [198]. Here, such assumptions correspond to the NN inductive bias, namely the CBI hypothesis that similar inputs have similar outputs. Given this hypothesis, the cases stored in the case base are no longer independent (grade 3 in French, in conjunction with this hypothesis, makes grade 3 in Spanish very likely).

The problem of interdependence cannot be taken into account as long as an estimation disregards the similarity between the instances stored in the memory (cf. Section 4.5.3), as do all the estimations presented so far. Still, the aggregation operator $\oplus$ in the possibilistic approach provides a means for alleviating the problem. With $\oplus = \max$, for example, frequency does not count at all and one obtains $\delta_{s_0}(3) = 1/3 < 3/4 = \delta_{s_0}(10)$. The probabilistic sum $\oplus_P$ brings about a reinforcement effect but still yields $\delta_{s_0}(3) = 0.7 < 3/4 = \delta_{s_0}(10)$, a result that appears quite reasonable.

A second problem related to scenario S2 is that of ambiguity. Particularly, the probabilistic approach yields a bimodal distribution $p_{s_0}$, and the same is also true for most aggregation operators in the possibilistic approach. For example, (5.14) with $\oplus = \oplus_P$ (and $\otimes = \otimes_P$) yields $\delta_{s_0}(3) > \delta_{s_0}(7) < \delta_{s_0}(10)$. This result is not intuitive, for one might hardly judge an intermediate grade less possible than two extreme grades. To solve this problem, $\delta_{s_0}$ can be replaced by its convex hull

$$r \mapsto \min \left\{ \max_{r' \leq r} \delta_{s_0}(r'), \max_{r' \geq r} \delta_{s_0}(r') \right\}. \tag{5.31}$$

In our example, this leads to the following distribution:

| $r$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_{s_0}(r)$ | 0 | 0.3 | 0.53 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.75 |

Of course, this prediction is still ambiguous in the sense that is supports several grades by means of high degrees of possibility. This is not a defect, however, but rather an adequate representation of the ambiguity which is indeed present in the situation associated with scenario S2.

The modification (5.31) of $\delta_{s_0}$ should not be considered ad-hoc. Rather, the convexity requirement can be thought of as a possibility-qualifying rule that complements the case-based justification of possibility degrees: The more possible two outputs are, the more possible is any outcome in-between. This type of background knowledge and the associated constraints can be met more easily in the possibilistic approach than in the probabilistic one. In fact, the incorporation of background information is hardly compatible with non-parametric density estimation.

In summary, the example has shown the following advantages of the possibilistic approach: Firstly, the interpretation of aggregated weights in terms of degrees of evidential support is often less critical than the interpretation in terms of degrees of probability. Secondly, a possibility distribution can represent ignorance. Thirdly, the use of aggregation operators other than the arithmetic sum can be useful. Fourthly, the possibilistic approach is more flexible and allows for incorporating constraints or background knowledge.

### 5.3.8 Complexity issues

A straightforward implementation of the prediction (5.13) has a running time which is linear in the size $|\mathcal{M}|$ of the memory and the number $|\mathcal{R}|$ of outcomes (resp. a discretization thereof). In this respect, it is hence completely comparable to other case-based learning methods.

In order to reduce the computational complexity, instance-based approaches take advantage of the fact that a prediction is already determined by the nearest neighbors of the query instance. Thus, the consideration of each sample instance is actually not necessary, and efficiency can be gained by means of fast algorithms for finding nearest neighbors [154, 411, 222]. Such algorithms employ efficient similarity-based indexing techniques and corresponding data structures in order to find the relevant instances quickly.

The same idea can be applied in connection with the possibilistic approach. In fact, a possibility degree $\delta_{s_0}(r)$ is completely determined by the neighborhood of the case $\langle s_0, r \rangle$, that is the sample instances $\langle s_\imath, r_\imath \rangle$ satisfying $\sigma_{\mathcal{S}}(s_\imath, s_0) > 0$ and $\sigma_{\mathcal{R}}(r_\imath, r) > 0$. As can be seen, apart from minor differences, the possibilistic method is quite comparable to other instance-based methods from a complexity point of view. One such difference concerns the relevant sample instances. In the $k$NN approach, the number of relevant instances in always $k$, but the (degree of) relevance of an instance may change when modifying the case base. As opposed to this, the degree of relevance of a neighboring instance is fixed in the possibilistic approach, but the number of relevant instances can change.

Let us finally mention that efficiency can also be gained if the complete possibility distribution $\delta_{s_0}$ is not needed. In fact, quite often one will only be interested in those outcomes having a high degree of possibility. For example, one might be interested in a fixed number of maximally supported outcomes, or in those outcomes whose support exceeds a given possibility threshold. In such cases, the computation of $\delta_{s_0}(r)$ can be omitted (or broken off) for certain outputs $r$.

## 5.4 Extensions of the basic model

The previous section has introduced the main principles of the possibilistic approach to case-based inference (subsequently, for the sake of brevity, sometimes referred to as PoCBI). In this regard, the close connection to fuzzy rule-based reasoning was especially emphasized. Besides, we highlighted the fact that possibilistic CBI can be considered as an alternative approach to NN estimation. This section presents some extensions of the basic model making PoCBI even more powerful and practically useful.

### 5.4.1 Dealing with incomplete information

The problem of dealing with incomplete information such as missing attribute values in an important issue in case-based reasoning and machine learning [88, 305]. For example, suppose that the specification of the new query $s_0$ is incomplete, and let $S_0 \subseteq \mathcal{S}$ denote the inputs compatible with the description of $s_0$. Moreover, recall the lower support-bound semantics of the possibilistic approach to CBI. The following generalization of (5.14) is in accordance with these semantics:

$$\begin{aligned} \delta_{s_0}(r) &\overset{\mathrm{df}}{=} \inf_{s \in S_0} \delta_s(r) = \\ &= \inf_{s \in S_0} \bigoplus_{1 \leq \imath \leq n} \sigma_{\mathcal{S}}(s, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath). \end{aligned} \tag{5.32}$$

Indeed, each potential candidate $s \in S_0$ gives rise to a lower bound according to (5.14), and without additional knowledge we can guarantee but the smallest of these bounds to be valid. This is in agreement with the idea of *guaranteed possibility* (cf. Section 5.1.2). The simplicity of handling incomplete information in a coherent (namely possibilistic) way is clearly a strong point of possibilistic CBI. Notice that the computation of the lower bound in (5.32) is in line with the handling of missing attribute values in the IB1 algorithm (cf. Section 2.2.2), where these values are assumed to be maximally different from the comparative value. Yet, the possibilistic solution appears more appealing since it avoids any default assumption. Indeed, inferring what is *possible* seems to be a reasonable way of dealing with missing attribute values and for handling incomplete and uncertain information in a coherent way.

Example 5.2. Reconsider Example 5.1 and suppose that we are interested in, say, the price of a car whose horsepower is between 90 and 110. This amounts to predicting the outcome of an income $s_0$, in which the attributes are incompletely specified. Fig. 5.1 shows the prediction obtained for the max–min version of (5.32) for this example. □

More generally, imprecise knowledge about $s_0$ can be modeled in the form of a possibility distribution $\pi$ on $\mathcal{S}$, where $\pi(s)$ corresponds to the degree of plausibility that $s_0 = s$. A graded modeling of this kind is useful, e.g., if some attributes are specified in a linguistic way. It suggests the following generalization of (5.32):

$$\delta_{s_0}(r) \overset{\mathrm{df}}{=} \inf_{s \in \mathcal{S}} \left( \pi(s) \rightsquigarrow \delta_s(r) \right), \tag{5.33}$$

where $\rightsquigarrow$ is a generalized implication operator that is reasonably chosen as the Gödel implication [134]:

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ \beta & \text{if} \quad \alpha > \beta \end{cases} .$$

From a logical point of view, (5.33) specifies the extent to which *the output $r$ is supported by all plausible candidates for* $s_0$. Notice that the distributions $\delta_s$ and $\pi$ in (5.32) have different semantics and express degrees of confirmation and plausibility, respectively (cf. Section 5.1). Particularly, $\pi$ is assumed to be normalized, i.e., there is at least one input $s$ with $\pi(s) = 1$. One obviously recovers (5.32) from (5.33) for the special case where $\pi$ is a $\{0,1\}$-valued possibility distribution $\pi = \mathbb{I}_{S_0}$ and hence corresponds to a crisp subset $S_0 \subseteq \mathcal{S}$.

Similar generalizations can also be realized for coping with incompletely specified examples. Let the $\imath$-th case in the memory be characterized by the set $S_\imath \times R_\imath \subseteq \mathcal{S} \times \mathcal{R}$. Then, (5.14) becomes

$$\delta_{s_0}(r) \overset{\text{df}}{=} \bigoplus_{1 \leq \imath \leq n} \inf_{\langle s', r' \rangle \in S_\imath \times R_\imath} \sigma_{\mathcal{S}}(s_0, s') \otimes \sigma_{\mathcal{R}}(r, r'),$$

which is in accordance with (5.32). Moreover, we obtain

$$\delta_{s_0}(r) \overset{\text{df}}{=} \bigoplus_{1 \leq \imath \leq n} \inf_{\langle s', r' \rangle \in \mathcal{S} \times \mathcal{R}} \max \left\{ \sigma_{\mathcal{S}}(s_0, s') \otimes \sigma_{\mathcal{R}}(r, r'), 1 - \pi_\imath(s', r') \right\}$$

if the $\imath$-th case is characterized by means of a possibility distribution $\pi_\imath$ on $\mathcal{S} \times \mathcal{R}$ rather than by a crisp set $S_\imath \times R_\imath$. Note that this expression can be combined with (5.33) in order to handle incomplete specifications of both, the sample cases and the new query. Moreover, notice that the distribution $\delta_{s_0}$ will generally remain unaffected if an example is completely unspecified ($\pi_\imath \equiv 1$), which is clearly a reasonable property.

Interestingly enough, the above generalization does not only allow for dealing with incomplete (fuzzy) cases. It also suggests to lump together several (similar) cases stored in the memory. The idea, then, is to replace these cases by one "fuzzy case", the attributes of which are given by the disjunction of the attribute values of the individual cases. On the one hand, this procedure might improve efficiency, especially if the memory of cases is very large. On the other hand, some information might be lost when basing a prediction on one or several fuzzy cases: In fact, it is not difficult to show that the support $\delta_{s_0}(r)$ of a (hypothetical) case $\langle s_0, r \rangle$ derived from a set of observed cases can be larger (but not smaller) than the support obtained from the fuzzy case which combines the original observations. Nevertheless, the more similar the combined observations are, the better the approximation becomes. Of course, instead of replacing a set of cases by a fuzzy case, one might also think of simply selecting one of these cases which is prototypical of this set.[18]

[18] This is in line with the idea of generating prototypes by merging training samples – and thus reducing the size of the training set – which has been proposed in the context of NN classification [62].

### 5.4.2 Discounting noisy and atypical instances

Since case-based prediction and instance-based learning are quite sensitive to noisy instances, it is reasonable to discard those instances [5]. By noise one generally means incorrect attribute value information, concerning either the descriptive part $s$ of a case or the outcome $r$ (or both). However, the problem of noise is also closely related to the "typicality" of a case. A typical case is representative of its neighbors, whereas an exceptional (though not incorrect) case has an outcome quite different from the outputs of neighboring cases [419].

Recall that each case $\langle s_\imath, r_\imath \rangle \in \mathcal{M}$ is extrapolated by placing the support function or, say, "possibilistic kernel" (5.13) around the point $\langle s_\imath, r_\imath \rangle \in \mathcal{S} \times \mathcal{R}$, just like a density (kernel) function is centered around each observation in kernel-based density estimation. Of course, the less representative (i.e., noisy or exceptional) a case is of its neighborhood, the smaller the extent of extrapolation should be.

A simple learning mechanism that adapts the extent of extrapolation of stored cases can be realized by means of a slight generalization of the kernel function (5.13):

$$\delta^{\imath}_{s_0} : r \mapsto m_\imath\big(\sigma_{\mathcal{S}}(s_0, s_\imath)\big) \; \otimes \; \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.34}$$

Here, $m_\imath : [0,1] \longrightarrow [0,1]$ is a monotone increasing modifier function with $m_\imath(1) = 1$. This function allows for discounting atypical cases. Roughly speaking, $m_\imath$ adapts the similarity between the instance $s_\imath$ and its neighbors. For example, $s_\imath$ is made completely dissimilar to all other instances by letting $(m_\imath | [0,1[) \equiv 0$. Replacing $\sigma_{\mathcal{S}}$ by the modified measure $m_\imath \circ \sigma_{\mathcal{S}}$ is closely related to the idea of local distance measures in NN algorithms.

Suppose that a new observation $s_0$ with output $r_0$ has been made, and consider a stored case $\langle s_\imath, r_\imath \rangle$. Should this case be discounted in the light of the new observation? The fact that $\langle s_\imath, r_\imath \rangle$ supports an outcome different from the observed output $r_0$ need not necessarily be a flaw. In fact, recall that $s_0 \in C_{r_0}$ does not exclude that $s_0 \in C_r$ for some $r \neq r_0$. In other words, neither the non-support of the observed nor the support of a different outcome can actually be punished. However, what can be punished is the disqualification of the output $r_0$ as expressed by the upper possibility model (5.19). Thus, it is reasonable to require that the degree of disqualification induced by $\langle s_\imath, r_\imath \rangle$ is limited:

$$1 - m_\imath(\sigma_{\mathcal{S}}(s_0, s_\imath)) \otimes \sigma_{\mathcal{R}}(r_0, r_\imath) \; \geq \; \beta, \tag{5.35}$$

where $\beta \gg 0$ is a constant.

The constraint (5.35) suggests an update scheme in which a stored case $\langle s_\imath, r_\imath \rangle$ is (maybe) discounted every time a new observation $\langle s_0, r_0 \rangle$ is made: Let $\mathcal{F}$ denote a parameterized and completely ordered class of functions from which $m_\imath$ is chosen. An adaptation is then realized by

$$m_\imath \leftarrow \min \big\{ m_\imath, \sup\{ f \in \mathcal{F} \,|\, 1 - f(\sigma_{\mathcal{S}}(s_0, s_\imath)) \otimes \sigma_{\mathcal{R}}(r_0, r_\imath) \; \geq \; \beta \} \big\} . \tag{5.36}$$

The discounting of noisy and atypical instances through modifying possibilistic kernel functions appears natural and somewhat simpler than the method used in IB3 [5]. Firstly, possibilistic discounting is gradual, whereas an instance is either accepted or rejected (or is temporarily in-between) in IB3. Secondly, the question whether to discount an instance and to which extent is answered quite naturally in the possibilistic approach, where support is absolute and graded. In IB3, an instance is either punished or not, and the corresponding decision is based on a rule that appears reasonable but might still be considered ad-hoc ($s_\imath$ is discounted if $\Delta_{\mathcal{S}}(s_\imath, s_0)$ is smaller than or equal to the distance between $s_0$ and its closest *accepted* neighbor[19]).

The possibilistic adaptation scheme becomes rather simple for the special case $\mathcal{S} = \mathfrak{R}^l$, $\mathcal{R} = \{0, 1\}$ and $m_\imath = \mathbb{I}_{]\gamma_\imath, 1]}$, where $0 \leq \gamma_\imath < 1$. If $\sigma_{\mathcal{S}}$ is a strictly decreasing function of Euclidean distance, then the support function (5.13) corresponds to a ball around $s_\imath$: $\delta^\imath_{s_0}(r) = 1$ if $r = r_\imath$ and $s_0$ is located inside that ball and $\delta^\imath_{s_0}(r) = 0$ otherwise. The parameter $\gamma_\imath$ is chosen as large as possible, but such that the support function does not cover any observed input $s_\jmath$ with $r_\jmath \neq r_\imath$, that is $\gamma_\imath \leq |s_\imath - s_\jmath|$ holds true for all of those $s_\jmath$. Fig. 5.2 gives an illustration for $l = 2$.

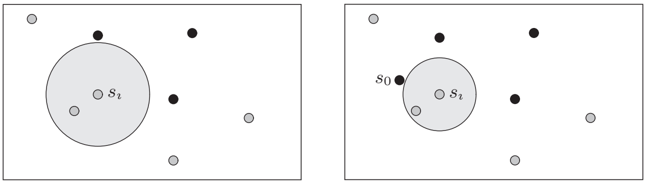


**Fig. 5.2.** Left: The large circle corresponds to the support function (possibilistic kernel) centered around $s_\imath$ and marks the extrapolation of outcome $r_\imath$. Right: The support function is updated after observing a new instance which has a different outcome $r_0 \neq r_\imath$ and hence must not be supported.

This special case, that we shall subsequently refer to as PossIBL, is a useful point of departure for investigating theoretical properties of the possibilistic approach in the context of concept learning. In [11], some convergence properties of IB1 have been shown for a special setup which makes statistical assumptions about the generation of training data and geometrical assumptions on a concept $C_1$ to be learned. For PossIBL, one can prove similar properties under the same assumptions. More specifically, let $l = 2$, $\mathcal{S} = [0, 1] \times [0, 1]$ (the results can be generalized to any dimension $l > 2$ and any bounded region $\mathcal{S} \subset \mathfrak{R}^l$) and consider a concept $C_1 \subseteq \mathcal{S}$. For the special case above, the PossIBL approximation of $C_1$ is then given by

$$C_1^{est} = \bigcup_{\langle s_\imath, 1\rangle \in \mathcal{M}} \mathfrak{B}_{\rho(s_\imath)}(s_\imath), \tag{5.37}$$

[19] Auxiliary rules are used if $s_0$ does not have an accepted neighbor.

where $\mathfrak{B}_d(s_\imath) = \{s \in \mathcal{S} \,|\, |s - s_\imath| < d\}$ is the (open) $d$-ball around $s_\imath$ and

$$\rho(s_\imath) = \min \left\{ |s_\jmath - s_\imath| \,|\, \langle s_\jmath, r_\jmath \rangle \in \mathcal{M},\, r_\jmath \neq r_\imath \right\}. \tag{5.38}$$

Moreover, the approximation of $C_0 = \mathcal{S} \setminus C_1$ is given by

$$C_0^{est} = \bigcup_{\langle s_\imath, 0 \rangle \in \mathcal{M}} \mathfrak{B}_{\rho(s_\imath)}(s_\imath). \tag{5.39}$$

It is readily verified that $C_0^{est} \cap C_1^{est} = \emptyset$. However, $C_0^{est} \cup C_1^{est} = \mathcal{S}$ does not necessarily hold true. Thus, one may have $\delta_{s_0} \equiv 0$ for some instances $s_0 \in \mathcal{S}$ (which are then classified at random). Consequently, an approximation of concept $C_1$ should actually be represented by the tuple $(C_0^{est}, C_1^{est})$ which divides instances $s_0 \in \mathcal{S}$ into three groups: Those which (supposedly) belong to $C_1$ ($\delta_{s_0}(0) = 0, \delta_{s_0}(1) = 1$), those which do not ($\delta_{s_0}(0) = 1, \delta_{s_0}(1) = 0$), and those for which no evidence is available so far ($\delta_{s_0} \equiv 0$).

Now, a first desirable property is the convergence of the concept approximation, that is the convergence of $C_0^{est}$ and $C_1^{est}$ toward $C_0$ and $C_1$, respectively. In this context, however, the property of convergence itself has to be weakened since exact convergence cannot be achieved due to the fact that an NN classifier cannot guarantee the avoidance of wrong decisions at the boundary of a concept. Moreover, some assumptions on the generation of samples and on the geometry of the concept $C_1$ have to be made. Here, we make the same assumptions as in [11]: Instances are generated randomly and independently according to a fixed probability measure $\mu$ over $\mathcal{S}$. Furthermore, $C_1$ is a concept having a *nice* boundary, which is the union of a finite number of closed (hyper-)curves of finite size.

We employ the following notation: The $\varepsilon$-neighborhood of $C_1$ is the set

$$C_1^+(\varepsilon) \overset{\text{df}}{=} \{s \in \mathcal{S} \,|\, \mathfrak{B}_\varepsilon(s) \cap C_1 \neq \emptyset\},$$

and the $\varepsilon$-core of $C_1$ is defined by

$$C_1^-(\varepsilon) \overset{\text{df}}{=} \{s \in \mathcal{S} \,|\, \mathfrak{B}_\varepsilon(s) \subseteq C_1\}.$$

A set $A \subseteq \mathcal{S}$ is called an $(\varepsilon, \gamma)$-approximation of $C_1$ if there is a (measurable) set $N \subseteq \mathcal{S}$ with $\mu(N) \leq \gamma$ and such that

$$(C_1^-(\varepsilon) \setminus N) \;\subseteq\; (A \setminus N) \;\subseteq\; (C_1^+(\varepsilon) \setminus N).$$

Finally, let $C_{1,n}^{est}$ and $C_{0,n}^{est}$ denote, respectively, the possibilistic concept approximations (5.37) and (5.39) for $|\mathcal{M}| = n$, i.e., after $n$ observations have been made.

**Lemma 5.3.** The equalities

$$C_1^-(\varepsilon) = \mathcal{S} \setminus C_0^+(\varepsilon) \quad \text{and} \quad C_0^-(\varepsilon) = \mathcal{S} \setminus C_1^+(\varepsilon)$$

hold true for all $0 < \varepsilon < 1$. □

**Proof.** For $s \in C_1^-(\varepsilon)$ we have $\mathfrak{B}_\varepsilon(s) \subseteq C_1$, which means that $|s - s_1| < \varepsilon$ implies $s_1 \in C_1$. Consequently, there is no $s_0 \in C_0$ such that $|s - s_0| < \varepsilon$ and, hence, $s \notin C_0^+(\varepsilon)$. Now, suppose $s \in \mathcal{S} \setminus C_0^+(\varepsilon)$. Thus, there is no $s_0 \in C_0$ such that $|s - s_0| < \varepsilon$, which means that $|s - s_1| < \varepsilon$ implies $s_1 \in C_1$ and, hence, $s \in C_1^-(\varepsilon)$. The second equality is shown in the same way. □

**Theorem 5.4.** Let $C_1 \subseteq \mathcal{S}$ and $0 < \varepsilon, \gamma, d < 1$. There is an integer $n_0$ such that the following holds true with probability at least $1 - d$: The possibilistic concept approximation $C_{1,n}^{est}$ is a $(2\varepsilon, \gamma)$-approximation of $C_1$ and $C_{0,n}^{est}$ is a $(2\varepsilon, \gamma)$-approximation of $C_0$ for all $n > n_0$. □

**Proof.** Let $N$ denote the set of instances $s \in \mathcal{S}$ for which no $s_\imath \in \mathcal{M}^\downarrow$ exists such that $|s - s_\imath| < \varepsilon$. In [11], the following lemma has been shown: $\mu(N) \leq \gamma$ holds true with probability $1 - d$ whenever

$$n > \lceil n_0 = \sqrt{2}/\varepsilon\rceil^2/\gamma^2 \cdot \ln\left(\lceil\sqrt{2}/\varepsilon\rceil^2/d\right). \tag{5.40}$$

Subsequently, we ignore the set $N$, that is we formally replace $\mathcal{S}$ by $\mathcal{S} \setminus N$, $C_1$ by $C_1 \setminus N$ and $C_0$ by $C_0 \setminus N$. Thus, the following holds true by definition: For each $s \in \mathcal{S}$ there is an instance $s_\imath \in \mathcal{M}^\downarrow$ such that $|s - s_\imath| < \varepsilon$.

Now, consider any instance $s \in C_1^-(2\varepsilon)$. We have to show that $s \in C_{1,n}^{est}$. Let $s_\imath \in \mathcal{M}^\downarrow$ be an instance such that $|s - s_\imath| < \varepsilon$. For this instance we have $s_\imath \in \mathfrak{B}_\varepsilon(s) \subseteq C_1$, which means that $s_\imath$ belongs to $C_1$. Furthermore, $\mathfrak{B}_\varepsilon(s_\imath) \subseteq \mathfrak{B}_{2\varepsilon}(s) \subseteq C_1$ and, hence, $\rho(s_\imath) \geq \varepsilon$ for the value in (5.38). This implies that $s \in \mathfrak{B}_{\rho(s_\imath)}(s_\imath)$ and, therefore, $s \in C_{1,n}^{est}$. Thus, we have shown that $C_1^-(2\varepsilon) \subseteq C_{1,n}^{est}$.

Since the same arguments apply to $C_0$, the property $C_0^-(2\varepsilon) \subseteq C_{0,n}^{est}$ can be shown in an analogous way. Thus, using Lemma 5.3,

$$C_{1,n}^{est} \subseteq \mathcal{S} \setminus C_{0,n}^{est} \subseteq \mathcal{S} \setminus C_0^-(2\varepsilon) = C_1^+(2\varepsilon).$$

Likewise, one shows that $C_{0,n}^{est} \subseteq C_0^+(2\varepsilon)$. □

Roughly speaking, Theorem 5.4 guarantees that the $2\varepsilon$-core of both, $C_0$ and $C_1$ is classified correctly (with high probability) if the memory $\mathcal{M}$ is large enough. In other words, classification errors can only occur in the boundary region. For being able to quantify the probability of an error, it is necessary to put restrictions on the size of that boundary region and on the probability distribution $\mu$. Thus, let $\mathcal{C}$ denote the class of concepts $C_1 \subseteq \mathcal{S}$ that can be represented as the union of a finite set of regions bounded by closed curves with total length of at most $L$ [11]. Moreover, let $\mathfrak{P}_\beta$ denote the class of probability distributions $\mu$ over $\mathcal{S}$ such that $\mu(A) \leq \mu_L(A) \cdot \beta$ for all Borel-subsets $A \subseteq \mathcal{S}$, where $\mu_L$ is the Lebesgue measure and $\beta > 0$.

**Theorem 5.5.** The concept class $\mathcal{C}$ is polynomially learnable with respect to $\mathfrak{P}_\beta$ by means of the possibilistic concept approximation $(C_0^{est}, C_1^{est})$. □

**Proof.** If $C_1 \in \mathcal{C}$, then the size of the region $C_1^+(2\varepsilon) \setminus C_1^-(2\varepsilon)$ is bounded by $4\,\varepsilon L$. Consequently, the probability of that area is at most $\alpha = 4\,\varepsilon L \beta$. Since a classification error can only occur either in this region or in the set $N$ as defined in Theorem 5.4 and the probability of $N$ is at most $\gamma$, the probability of a classification error is bounded by $\alpha + \gamma$. Now, fix the parameters $\gamma$ and $\varepsilon$ as follows: $\gamma = e/2$, $\varepsilon = e/(8L\beta)$. By substituting these parameters into (5.40) one finds that the required sample size $n$ is polynomial in $1/e$ and $1/d$. In summary, the following holds true for any $0 < e, d < 1$, $C_1 \in \mathcal{C}$, and $\mu \in \mathfrak{P}_\beta$: If more than $n(1/e, 1/d)$ examples are presented, where $n$ is a polynomial function of $1/e$ and $1/d$, then, with probability $1 - d$, the possibilistic concept approximation has a classification error of at most $e$. This is precisely the claim of the theorem. □

### 5.4.3 From instances to rules

As already mentioned in previous chapters, selecting appropriate cases to be stored in the memory $\mathcal{M}$ is an important issue in case-based reasoning and instance-based learning that has a strong influence on performance. Especially reducing the size of the memory is often necessary in order to maintain the efficiency of the system. The basic idea is to remove cases which are actually not necessary to achieve good predictive performance. For example, consider the problem of concept learning and imagine a concept having the form of a circle in some (two-dimensional) instance space. To classify inner points correctly by means of the $k$NN rule it might then be sufficient to store positive examples of that concept near the boundary.

In connection with PossIBL, where support is absolute rather than relative, deleting cases from the memory might produce “holes” in the concept description. An interesting alternative, which allows one to reduce the size of the memory and, at the same time, to fill “holes” in the concept description by interpolation, is based on the idea of merging cases and of generalizing cases into rules. This idea appears particularly reasonable in light of the close relation between PoCBI and fuzzy rule-based reasoning. More precisely, each observation can be interpreted as a fuzzy rule, namely as an instance of a fuzzy meta-rule suggesting that similar inputs (possibly) have similar outputs.

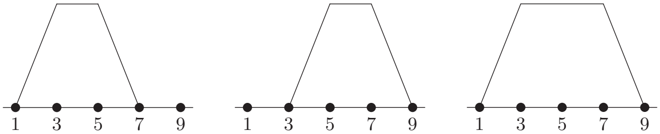


**Fig. 5.3.** Possibility distributions induced by two cases (left, middle) and the distribution associated with the summarizing fuzzy rule (right).

To illustrate this idea of a one-to-one correspondence between rules and cases, let $\mathcal{S} = \mathfrak{R}$, $\mathcal{R} = \{0,1\}$ and suppose that two inputs $s_1 = 4$ and $s_2 = 6$ with $r_1 = r_2 = 0$ have been observed. The possibilistic kernels (5.13) induced by these cases are shown in Fig. 5.3. The first case is equivalent to the fuzzy rule "If $s_0$ is approximately 4 then $r = 0$" if the fuzzy set "approximately 4" is modeled by the possibility distribution $\delta^1_{s_0}$ (the individual support function (5.13)). The rules associated with the two cases can be merged into one rule, say, "If $s_0$ is about 5 then $r = 0$", where the fuzzy set "about 5" is modeled by the pointwise maximum, $\delta^1_{s_0} \vee \delta^2_{s_0}$, of $\delta^1_{s_0}$ and $\delta^2_{s_0}$ (Fig. 5.3, right).

The above procedure is closely related to several other techniques that have been proposed in connection with IBL. Viewing cases as maximally specific rules and the idea of generalizing cases into rules has been put forward in [89, 90]. The method proposed in [327] generalizes cases by placing rectangles of different size around them. A new instance is then labeled by the nearest rectangle rather than by the nearest case. This is very similar to our approach, where rectangles are replaced by possibility distributions. Relations also exist with the idea of merging nearest neighbors of the same output (class label in classification), thereby generating new (pseudo-sample) prototypes [62].[20] In our example, the point 5 may be regarded as a pseudo-instance replacing 4 and 6 (and also endowed with a modified support function).

In the example in Fig. 5.3, the summarizing rule is exactly equivalent to the conjunction of the two individual rules. Of course, by weakening the requirement of equivalence, the merging procedure might also incorporate concepts of approximation and interpolation. For example, suppose $s_2 = 8$ rather than $s_2 = 6$. The replacement of $\delta^1_{s_0} \vee \delta^2_{s_0}$ by its convex hull $\delta : s \mapsto \max\{\delta^1_{s_0}(s), \delta^2_{s_0}(s), \mathbb{I}_{[5,7]}\}$ then goes beyond a simple combination since $\delta$ is larger than the pointwise maximum of $\delta^1_{s_0}$ and $\delta^2_{s_0}$ (e.g. $\delta^1_{s_0}(6) = \delta^2_{s_0}(6) = 0.5 < 1 = \delta(6)$). This kind of possibilistic induction can be reasonable and often allows for incorporating background knowledge. Particularly, replacing a possibilistic estimation $\delta_{s_0}$ by its convex hull is advised whenever a multimodal distribution does not make sense (as in our example in Section 5.3.7) or if the relation of observable cases (cf. page 22) is even known to satisfy a convexity constraint of the form

$$s \in C_r \cap C_{r''} \Rightarrow s \in C_{r'} \tag{5.41}$$

for all $r < r' < r''$.

As can be seen, the extensions discussed here basically suggest a system that maintains an optimal rule base rather than an optimal case base, including the combination and adaptation of rules. These extensions are well-suited to the discounting of cases discussed in Section 5.4.2. Indeed, deriving one rule from several cases (or other rules) can be accomplished by replacing the latter by a pseudo-case and defining an appropriate modifier function $m$ for that pseudo-instance.

[20] Compare also with the idea of "fuzzy cases" discussed at the end of Section 5.4.1.

### 5.4.4 Modified possibility rules

The basic model of possibilistic CBI introduced in Section 5.2 can be rendered more flexible by making use of (linguistic) modifiers [413] in (5.7), i.e., non-decreasing functions $m_1, m_2 : [0,1] \longrightarrow [0,1]$. This leads to possibility rules $m_1 \circ A \overset{m_2}{\rightsquigarrow} B$ with associated distributions

$$\delta_{s_0}(r) = \max_{1 \leq \imath \leq n} m_2 \left( \min \left\{ m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath) \right\} \right) , \tag{5.42}$$

or, when using generalized logical operators as suggested in Section 5.3,

$$\delta_{s_0}(r) = \bigoplus_{1 \leq \imath \leq n} m_2 \left( m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)) \otimes \sigma_{\mathcal{R}}(r, r_\imath) \right) .$$

Both modifiers in (5.42) control the extent to which a sample case is extrapolated, i.e., the extent to which other (hypothetical) cases are supported by an observation. The larger (in the sense of the partial order of functions on $[0,1]$) $m_1$ and $m_2$ are, the stronger (in the sense of asserted possibility degrees) a case $\langle s_\imath, r_\imath \rangle$ is extrapolated.

The modification (5.42) can be interpreted in different ways. Let us first consider the function $m_1$. In connection with the linguistic modeling of fuzzy concepts, modifiers such as $x \mapsto x^2$ or $x \mapsto \sqrt{x}$ are utilized for depicting the effect of linguistic hedges such as "very" or "almost" [413]. Applying the modifier $m_1$ defined by the mapping $x \mapsto x^2$ might thus be seen as replacing the original hypothesis that "similar inputs (possibly) induce similar outcomes" by the weaker assumption that only "*very* similar situations (possibly) induce similar outcomes." Thus, one interpretation of (5.42) is that of adapting the CBI hypothesis and, hence, the inference mechanism (but of maintaining the similarity measures): "The more two inputs are $m_1$-similar in the sense of $\sigma_{\mathcal{S}}$, the more possible it is that the respective results are (at least) similar in the sense of $\sigma_{\mathcal{R}}$."

According to a second interpretation the similarity measure $\sigma_{\mathcal{S}}$ is replaced by the measure $\sigma'_{\mathcal{S}} = m_1 \circ \sigma_{\mathcal{S}}$ in such a way that the CBI hypothesis applies in its original form:[21] "The more two inputs are similar in the sense of $\sigma'_{\mathcal{S}}$, the more possible it is that the respective results are (at least) similar in the sense of $\sigma_{\mathcal{R}}$." Roughly speaking, not the hypothesis is adapted to similarity, but similarity to the hypothesis. The extreme example $m_1 = \mathbb{I}_{\{1\}}$, indicating that the CBI hypothesis is not satisfied at all, again reveals that a similarity measure which is reasonable in the sense of inducing an appropriate extrapolation of observations does not necessarily appear natural. Indeed, interpreting $\sigma'_{\mathcal{S}} = m_1 \circ \sigma_{\mathcal{S}}$ as an improved measure suggests that inputs are not comparable at all.

The modifier $m_2$ does not act on a similarity measure but on the possibility-qualifying part of a rule. It can be thought of as modifying the possibility distribution

[21] One has to be careful with this interpretation, since modified measures do not necessarily inherit all (mathematical) properties of the original relations.

$$(s,r) \mapsto \max_{1 \leq \imath \leq n} \min\{m_1(\sigma_{\mathcal{S}}(s,s_\imath)), \sigma_{\mathcal{R}}(r,r_\imath)\} \tag{5.43}$$

associated with the possibility rule $m_1 \circ A \rightharpoondown B$. In fact, it allows for modeling rules of the form "for $m_1$-similar inputs it is $m_2$-possible that the respective results are similar," where "$m_2$-possible" stands for expressions like "more or less possible." Linguistic hedges such as "more or less" basically bring about a discounting of the distribution (5.43) and, hence, of the rule $m_1 \circ A \rightharpoondown B$.

Discounting a possibility distribution $\delta$ can be accomplished in different ways. A simple approach which is also applicable within the framework of qualitative possibility theory (where similarity and possibility are measured on ordinal scales) is to modify $\delta$ into $\min\{1-\lambda, \delta\}$ [120]. The constant $\lambda$ plays the role of a discounting factor and defines an upper bound to the support that can be provided by an underlying (possibility) rule. Indeed, $\delta$ remains unchanged if $\lambda = 0$. As opposed to this, the original support expressed by $\delta$ is completely annulled if the discounting is maximal ($\lambda = 1$). By taking $m_2$ as the mapping $x \mapsto \min\{1-\lambda, x\}$, the distribution (5.42) becomes

$$\delta_{\mathcal{C}} : (s,r) \mapsto \max_{1 \leq \imath \leq n} \min\left\{1-\lambda, \min\left\{m_1(\sigma_{\mathcal{S}}(s,s_\imath)), \sigma_{\mathcal{R}}(r,r_\imath)\right\}\right\}. \tag{5.44}$$

Note that the similarity measure $\sigma_{\mathcal{R}}$ is not modified directly. Thus, it somehow determines the granularity of the extrapolation and, hence, the possibilistic approximation (5.44).

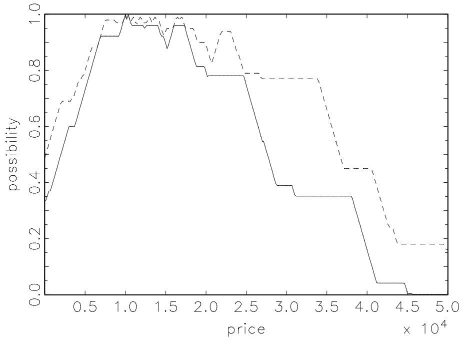


**Fig. 5.4.** Prediction (5.8) of the price of a car based on the original hypothesis (dashed line) and its modified version (5.44).

EXAMPLE 5.6. Reconsider Example 5.1 with the hypothesis that "it is completely possible that cars with *very* similar horsepower have similar prices." Applying the modifier $m_1 : x \mapsto x^2$ to the similarity relation $\sigma_{hp}$ and modeling the (non-)effect of "completely" by $\lambda = 0$, the prediction $\delta_{s_0}$ based on (5.44) yields the possibility distribution shown in Fig. 5.4. Compared to the prediction (5.8), the degree of possibility is smaller for most of the prices $r \in \mathcal{R}$. This is caused by the fact that the CBI hypothesis is now modeled in a more cautious way. □

### 5.4.5 Combination of several rules

Rather than making use of a single possibility rule, the CBI hypothesis can be expressed by means of a combination (conjunction) of several rules. Suppose $m$ such rules to be specified. Denoting by $\delta^k_{s_0}$ the possibility distribution (5.8) induced by the $k$-th rule ($1 \leq k \leq m$), the overall prediction is then given by

$$\delta_{s_0}(r) = \delta^1_{s_0}(r) \vee \delta^2_{s_0}(r) \vee \ldots \vee \delta^m_{s_0}(r). \tag{5.45}$$

The *disjunctive* combination in (5.45) shows that an outcome can be supported by any observed case in connection with any rule. Notice that each rule might involve different similarity relations, or different modifications of basic relations. Within our framework, it seems particularly interesting to compose new measures from a set of elementary relations (associated with individual attributes) by means of fuzzy set-based modeling techniques.

Suppose, as in the Example 5.1, that an attribute–value representation is used in order to characterize cases. That is, let inputs correspond to vectors $s = (a_1, \ldots, a_L) \in \mathcal{S} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_L$, where $\mathcal{A}_\jmath$ denotes the domain of the $\jmath$-th attribute. Moreover, let $\sigma_\jmath$ be an elementary similarity relation defined over $\mathcal{A}_\jmath$. By making use of logical connectives, the antecedent part of a possibility rule can then be composed of these elementary measures or modified versions thereof. Restricting ourselves to the logical connective $\wedge$, we obtain rules of the form

$$m_{11}(\sigma_1(a_1, a'_1)) \wedge \ldots \wedge m_{1L}(\sigma_L(a_L, a'_L)) \overset{m_2}{\rightharpoondown} \sigma_{\mathcal{R}}(r, r'). \tag{5.46}$$

Such rules can also be expressed as $\sigma'_{\mathcal{S}} \overset{m_2}{\rightharpoondown} \sigma_{\mathcal{R}}$, where

$$\sigma'_{\mathcal{S}}(s, s') = \bigotimes_{1 \leq \jmath \leq L} m_{1\jmath}(\sigma_\jmath(a_\jmath, a'_\jmath)), \tag{5.47}$$

provided that the elementary similarity relations in (5.47) are commensurate.

Of course, the antecedent part in (5.46) can be generalized such that only some of the attributes are used, i.e., each rule can concern different attributes. Leaving the $\jmath$-th attribute out of account can be interpreted in two ways. Firstly, this attribute might be irrelevant for the similarity of inputs, which is adequately reflected by $m_{1\jmath} \equiv 1$. Secondly, the rule might be interpreted as expressing a ceteris paribus condition, i.e., it might be assumed implicitly that $a_\jmath = a'_\jmath$. In this case, $m_{1\jmath}$ should be defined as $m_{1\jmath}(1) = 1$ and $m_{1\jmath}(x) = 0$ for $0 \leq x < 1$.[22] For

[22] Besides, $\sigma_\jmath$ should be separating.

example, when saying that two cars with similar horsepower have similar prices, it might be taken for granted that both cars have the same type of aspiration (standard or turbo).

Suppose that $m$ possibility rules have been defined by using the same modifier $m_2$. Moreover, let $\sigma_{\mathcal{S}}^k$ $(1 \leq k \leq m)$ denote the (aggregated) measure (5.47) associated with the antecedent part of the $k$-th rule. Thus, the rules specify different conditions (in the form of conjunctions of similarity relations between attributes) which allow for drawing the same conclusion. The $m$ individual rules are then equivalent to one (aggregated) rule of the form $\sigma_{\mathcal{S}} \overset{m_2}{\rightharpoondown} \sigma_{\mathcal{R}}$, where

$$\sigma_{\mathcal{S}}(s, s') = \bigoplus_{1 \leq k \leq m} \sigma_{\mathcal{S}}^k(s, s').$$

That is, the antecedent part of the aggregated rule corresponds to the disjunction of the antecedent parts of the individual rules.

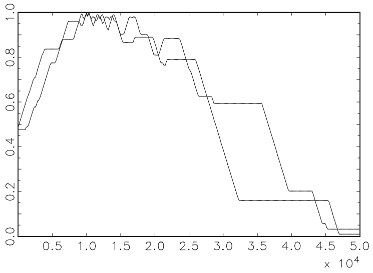


**Fig. 5.5.** Prediction (5.42) of the price of a car with horsepower 100, engine-size 110 and peak-rpm 5500, induced by two different rules.

Example 5.7. Reconsider Example 5.1 and let the following rules be given: (1) Cars with *very similar* horsepower possibly have similar prices. (2) Cars with *similar* engine-size and *approximately similar* peak-rpm (revolutions per minute) possibly have similar prices. Making use of the similarity measures $\sigma_{eng}(x, x') = \max\{1 - |x - x'|/100, 0\}$ and $\sigma_{rpm}(x, x') = \max\{1 - |x - x'|/1000, 0\}$, respectively, and modeling the effect of the linguistic hedge "approximately" by means of $x \mapsto \sqrt{x}$, the two rules yield the two predictions shown in Fig. 5.5. The overall prediction associated with the conjunction of the rules (i.e., the disjunction of the two premises) corresponds to the pointwise maximum of these distributions. □

Of course, different rules (5.46) will generally use different modifiers $m_2$. They should then be consistent in the sense that a strengthening of the antecedent part of a rule does not entail a reduction of extrapolation. Thus, consider two rules (5.46) modeled by means of modifiers $m^1_{1\jmath}, m^1_2$ and $m^2_{1\jmath}, m^2_2$ $(1 \leq \jmath \leq L)$, respectively. The first rule is obviously redundant with respect to the second one if

$$\forall\, 1 \leq \jmath \leq L \,:\, m^1_{1\jmath} \leq m^2_{1\jmath} \quad \text{and} \quad m^1_2 \leq m^2_2.$$

In fact, we then have $\delta^1_{s_0} \leq \delta^2_{s_0}$ for the possibility distributions induced by these two rules in connection with any observed case.

Consider the following rules as an example: (1) For cars with *similar* horsepower it is *completely* possible that the associated prices are similar. (2) For cars with *very similar* horsepower it is *more or less* possible that the associated prices are similar. This example reveals that redundancy always emerges in connection with somewhat conflicting rules (a stronger condition entails a weaker conclusion). Therefore, redundant rules should be avoided.

### 5.4.6 Locally restricted extrapolation

So far, the possibility rules which define a model of the CBI hypothesis have been used *globally* in the sense that they apply to all cases of the input-output space $\mathcal{S} \times \mathcal{R}$. Needless to say, the CBI hypothesis does not necessarily apply equally well to all parts of this space. That is to say, the degree of extrapolation of a case $\langle s, r \rangle$ that can be justified by the CBI hypothesis might depend on the region to which it belongs.

In the AUTOMOBILE DATABASE database (cf. Example 5.1), for instance, the variance of the price is smaller for cars with aspiration “turbo” than for cars with aspiration “standard” (even though the average price is higher for the former). Thus, the hypothesis that similar cars possibly have similar prices seems to apply better to turbo than to standard cars. Likewise, a statistical analysis suggests that the variation of the price is an increasing function of the size of cars. Again, the smaller a car is, the better the CBI hypothesis seems to apply (at least if the similarity of two lengths $x, x'$ is a function of $|x - x'|$). Consequently, the extrapolation of case-based information should be larger for small cars than for large cars.

In order to adapt the formalization of the CBI hypothesis one might think of defining different rules for different regions of the input space. Restricting the application of a rule to a certain (fuzzy) range of this space can be accomplished by means of a fuzzy partition $\mathcal{F}$ of $\mathcal{S}$. The condition part of a rule then appears in the form

$$F(s) \wedge F(s') \wedge m_1(\sigma_{\mathcal{S}}(s, s')), \tag{5.48}$$

where the fuzzy set $F \in \mathcal{F}$ is identified by its membership function $F : \mathcal{S} \longrightarrow [0, 1]$. The antecedent (5.48) can be associated with an extended possibility rule “the more both inputs are in $F$ *and* the more similar they are, the more possible it is that the related outcomes are similar.” This way, one might express, for instance, that "it is completely possible that small cars of similar size have similar prices" and "it is more or less possible that large cars of similar size have similar prices." The fuzzy set $F$ in (5.48) is then given by the set of small cars and large cars, respectively. Note that the attribute "aspiration" defines a crisp rather than a fuzzy partition.

On the basis of (5.48), the inference scheme (5.42) becomes

$$\begin{aligned}\delta_{s_0}(r) = \max_{1 \leq \imath \leq n} \min \big\{ F(s_0), F(s_\imath), \\ m_2 \left( \min \{ m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath) \} \right) \big\} .\end{aligned} \tag{5.49}$$

Note that $\delta_{s_0} \equiv 0$ as soon as $F(s) = 0$, thus expressing that a rule has no effect outside its region of applicability. Besides, it is worth mentioning that (5.49) is closely related to ideas of discounting as discussed in previous sections. This becomes especially apparent when writing (5.49) in the form

$$\delta_{s_0}(r) = \max_{1 \leq \imath \leq n} m_{2\imath}(x_\imath), \tag{5.50}$$

with $x_\imath = \min\{m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath)\}$ and $m_{2\imath} : x \mapsto \min\{F(s_0), F(s_\imath), m_2(x)\}$. In fact, (5.50) shows that the original support provided by the cases is discounted by means of the modifiers $m_{2\imath}$. As opposed to (5.44), however, this is not realized by using a constant factor $\lambda$. Rather, the discounting of a rule now depends on the inputs $s$ and $s_\imath$ to which it is applied.

### 5.4.7 Incorporation of background knowledge

Our fuzzy set-based framework is also well-suited for incorporating background knowledge of more general nature (i.e., not necessarily related to similarity). This becomes especially apparent if such knowledge is also expressed in terms of fuzzy rules. For instance, an expert might be willing to agree that "a price of slightly more than $40,000 for a car with horsepower of approximately 200 is completely possible." This can be formalized as a possibility rule $A \rightharpoondown B$, where $A$ and $B$ model the fuzzy sets of "approximately 200" and "slightly more than $40,000." Such a rule can simply be added to the rule base induced by the memory of cases (cf. Section 5.4.3), thereby supplementing the "empirical" evidence which comes from observed cases.

A special type of (rule-based) background knowledge can be obtained by specifying "fictitious cases". One might specify, for instance, a fictitious car by means of some attribute values (which can be uncertain or vague) and then ask an expert for a typical (or possible) price. The fictitious observation thus defined can principally be treated in the same way as an observed one. This type of reasoning provides a convenient way of filling up sparse memories. It is also interesting from a knowledge acquisition point of view. Indeed, from a user (expert) perspective it might appear less difficult to give some specific examples (e.g., by estimating prices of hypothetical cars) than to specify universally valid rules.

Apart from fuzzy rules, more general types of constraints can be used for expressing background knowledge. A nice example is the convexity constraint (5.41) according to which intermediary predictions are not less possible than more extreme ones. In order to satisfy such a constraint, a possibility distribution $\delta_{s_0}$ can simply be replaced by its convex hull (see (5.31) in Section 5.3.7).

## 5.5 Experimental studies

### 5.5.1 Preliminaries

This section presents some experimental studies providing evidence for the excellent practical performance of the possibilistic approach to case-based inference. More specifically, we shall focus on simple classification problems and investigate the PossIBL algorithm as introduced in Section 5.4.2. As in previous chapters, however, we would like to emphasize that our experiments are not meant as an exhaustive comparative study covering several competing learning algorithms – and showing that PossIBL is superior to all of its competitors. In fact, one should realize that the primary motivation underlying PossIBL (or, more generally, PoCBI) is not another $\varepsilon$-improvement in classification accuracy but rather the enrichment of instance-based learning (case-based reasoning) by concepts of possibilistic reasoning (though the latter does clearly not exclude the former). Besides, one should keep the following points in mind. Firstly, PossIBL has not been developed within a statistical framework. Thus, the type of problems for which PossIBL is most suitable (see the example in Section 5.3.7) is perhaps not represented in the best way by standard (public) data sets commonly used for testing performance. Secondly, an important aspect of the possibilistic approach is the one of *knowledge representation*. But this aspect is neglected if – as in experimental studies – only the correctness of the final decision (classification accuracy) counts, not the estimated distribution. Thirdly, regarding other IBL algorithms, a comparison might appear dubious since PossIBL – in its most general form – is an *extension* of IBL and hence covers specific algorithms such as $k$NN as special cases.

Due to these reasons, we have decided to apply a basic version of PossIBL to several data sets from the UCI repository[23] and to employ the $k$NN (resp. IB1) algorithm as a reference (we use $k$NN with $k = 1, 3, 5$ and the weighted 5NN rule with weight function (2.9)). Thus, we have refrained from tuning various degrees of freedom in order to optimize the performance of PossIBL (an exception is only the experimental study presented in Section 5.5.4). Instead, we have applied

[23] `http://www.ics.uci.edu/~mlearn`.

the learning scheme from Section 5.4.2 with the original max–min version (5.8). The function $m_\imath$ in (5.34) was defined as $t \mapsto \exp(-\gamma_\imath (1-t))$, where $\gamma_\imath \geq 0$ is the discounting rate of the $\imath$-th case. The constant $\beta$ in (5.35) was taken as 0.8.[24] In order to avoid difficulties due to the different handling of non-nominal class labels and the definition of similarity measures for non-numeric attributes, we have restricted ourselves to data sets for which all predictive attributes are numeric and for which the class label is defined on a nominal scale. The similarity $\sigma_{\mathcal{S}}$ is always defined as 1 minus the normalized Euclidean distance and the similarity $\sigma_{\mathcal{R}}$ is given by (5.11).

### 5.5.2 Classification accuracy

The experiments in this section were performed as follows: In a single simulation run, the data set is divided at random into a training set (the memory $\mathcal{M}$) and a test set, and the discounting rates $\gamma_\imath$ are adapted to the training set. A decision is then derived for each element of the test set by extrapolating the training set (but without adapting the discounting rates or expanding the memory any further), and the percentage of correct decisions is determined. Statistics are obtained by means of repeated simulation runs.

| Algorithm | mean | std. | min | max | 0.1–frac. | 0.9–frac. |
|---|---|---|---|---|---|---|
| PossIBL | 0.8776 | 0.0148 | 0.8215 | 0.9230 | 0.8584 | 0.8984 |
| 1NN | 0.7837 | 0.0161 | 0.7323 | 0.8369 | 0.7630 | 0.8030 |
| 3NN | 0.8117 | 0.0165 | 0.7630 | 0.8707 | 0.7907 | 0.8338 |
| 5NN | 0.8492 | 0.0155 | 0.8030 | 0.8923 | 0.8307 | 0.8707 |
| w5NN | 0.7864 | 0.0164 | 0.7294 | 0.8428 | 0.7655 | 0.8067 |

**Table 5.1.** Results for the Balance Scale Database (625 observations, 4 predictive attributes, three classes, training set of size 300, $1,000$ simulation runs).

| Algorithm | mean | std. | min | max | 0.1–frac. | 0.9–frac. |
|---|---|---|---|---|---|---|
| PossIBL | 0.9574 | 0.0204 | 0.8400 | 1.0000 | 0.9333 | 0.9733 |
| 1NN | 0.9492 | 0.0196 | 0.8400 | 1.0000 | 0.9200 | 0.9733 |
| 3NN | 0.9554 | 0.0175 | 0.8666 | 1.0000 | 0.9333 | 0.9733 |
| 5NN | 0.9586 | 0.0181 | 0.8533 | 1.0000 | 0.9333 | 0.9866 |
| w5NN | 0.9561 | 0.0187 | 0.8400 | 1.0000 | 0.9333 | 0.9733 |

**Table 5.2.** Results for the Iris Plant Database (150 observations, 4 predictive attributes, three classes, training set of size 75, $10,000$ simulation runs).

[24] Variations of this parameter had no significant influence.

| Algorithm | mean | std. | min | max | 0.1–frac. | 0.9–frac. |
|---|---|---|---|---|---|---|
| PossIBL | 0.6841 | 0.0419 | 0.5300 | 0.8400 | 0.6300 | 0.7400 |
| 1NN | 0.6870 | 0.0410 | 0.5200 | 0.8200 | 0.6300 | 0.7400 |
| 3NN | 0.6441 | 0.0421 | 0.4800 | 0.8100 | 0.5900 | 0.7000 |
| 5NN | 0.6277 | 0.0412 | 0.4800 | 0.7800 | 0.5700 | 0.6800 |
| w5NN | 0.6777 | 0.0414 | 0.5000 | 0.8300 | 0.6200 | 0.7300 |

**Table 5.3.** Results for the Glass Identification Database (214 observations, 9 predictive attributes, seven classes, training set of size 100, 10,000 simulation runs).

| Algorithm | mean | std. | min | max | 0.1–frac. | 0.9–frac. |
|---|---|---|---|---|---|---|
| PossIBL | 0.7096 | 0.0190 | 0.6421 | 0.7711 | 0.6868 | 0.7316 |
| 1NN | 0.6707 | 0.0199 | 0.6132 | 0.7289 | 0.6447 | 0.6947 |
| 3NN | 0.6999 | 0.0183 | 0.6447 | 0.7500 | 0.6763 | 0.7237 |
| 5NN | 0.7190 | 0.0183 | 0.6553 | 0.7684 | 0.6947 | 0.7421 |
| w5NN | 0.6948 | 0.0188 | 0.6421 | 0.7474 | 0.6684 | 0.7184 |

**Table 5.4.** Results for the Pima Indians Diabetes Database (768 observations, 8 predictive attributes, two classes, training set of size 380, 1,000 simulation runs).

| Algorithm | mean | std. | min | max | 0.1–frac. | 0.9–frac. |
|---|---|---|---|---|---|---|
| PossIBL | 0.7148 | 0.0409 | 0.5506 | 0.8652 | 0.6629 | 0.7640 |
| 1NN | 0.7163 | 0.0408 | 0.5843 | 0.8652 | 0.6629 | 0.7640 |
| 3NN | 0.6884 | 0.0407 | 0.5506 | 0.8315 | 0.6404 | 0.7416 |
| 5NN | 0.6940 | 0.0392 | 0.5730 | 0.8090 | 0.6404 | 0.7416 |
| w5NN | 0.7031 | 0.0404 | 0.5730 | 0.8315 | 0.6517 | 0.7528 |

**Table 5.5.** Results for the Wine Recognition Data (178 observations, 13 predictive attributes, three classes, training set of size 89, 1,000 simulation runs).

Results are summarized in Tables 5.5.2–5.5.2 by means of statistics for the percentage of correct classifications (mean, standard deviation, minimum, maximum, 0.1–fractile, 0.9–fractile). The experiments show that PossIBL achieves comparatively good results and is always among the best algorithms. Thus, it is valid to conclude that even a very basic version of PossIBL performs at least as well as the basic IBL (NN) algorithms. In other words, possibilistic IBL is in no way inferior to "standard" IBL as a basis for further improvements and sophisticated learning algorithms.

Due to the special setting of our experimental studies, especially the choice of max as an aggregation operator and the use of a $\{0, 1\}$-valued similarity measure over $\mathcal{R}$, one might wonder how to explain the different performance of PossIBL and the NN classifiers. In fact, in Section 5.3.6 it was argued that the possibilistic NN decision derived from (5.8) is actually equivalent to the 1NN rule when applying the maximum operator. It should hence be recalled that PossIBL, as employed in the above experiments, involves an adaptation of the (absolute) possibilistic support that comes from stored cases, which in essence is responsible for the differences.

A very interesting finding is the following: In the above examples, classification performance of the $k$NN algorithm is generally an increasing or a decreasing function of $k$. PossIBL, on the other hand, performs very well irrespective of the direction of that tendency, i.e., regardless of whether a smaller or a larger neighborhood should be called in. This can be taken as an indication of the robustness of the possibilistic approach.

### 5.5.3 Statistical assumptions and robustness

Let us elaborate a little more closely on the aspect of robustness. Above, it has been claimed that the possibilistic approach is more robust than other methods against violations of statistical assumptions of independence (see page 185). This is clearly true for the possibilistic estimation $\delta_{s_0}$ the informational content of which remains meaningful even if data is not independent. Here, we would like to provide experimental evidence for the supposition that the possibilistic approach can indeed be advantageous from both, an estimation and a decision making point of view, if the sample is not fully representative of the population.

The experimental setup is as follows: The instance space is defined by $\mathcal{S} = \mathfrak{R}$, the set of class labels is $\mathcal{R} = \{-1, +1\}$, the class probabilities are 1/2, the conditional probability density of the input $s$ given the outcome $r$ is normal with standard deviation 1 and mean $r$. In a single simulation run, a random sample of size $n = 20$ is generated, using class-probabilities of $1/2 - \alpha$ and $1/2 + \alpha$, respectively ($0 < \alpha \leq 1/2$). Based on the resulting training set, which is not "fully representative" in the sense of [78], predictions are derived for 10 new instances. These instances, however, are generated with the true class-probabilities of 1/2. For a fixed value $\alpha$ and a fixed prediction method, a misclassification rate $f(\alpha)$ is derived by averaging over 10,000 simulation runs.

Fig. 5.6 shows the misclassification rates for several methods. As was to be expected, $f(\cdot)$ is an increasing function of the sample bias $\alpha$. The best results are of course obtained if the class-probabilities of the training set and the test set coincide, that is for $\alpha = 0$. The figure also reveals that the sensitivity of the $k$NN classifier increases with $k$. On the one hand, it is true that a larger $k$ leads to better results for $\alpha$ close to 0. On the other hand, the performance decreases more quickly than for smaller $k$, and $k = 1$ is to be preferred for $\alpha$ close to $1/2$. This finding can also be grasped intuitively: The larger $k$, the more the $k$NN rule relies on frequency information, and the more it is affected if this information is misleading.

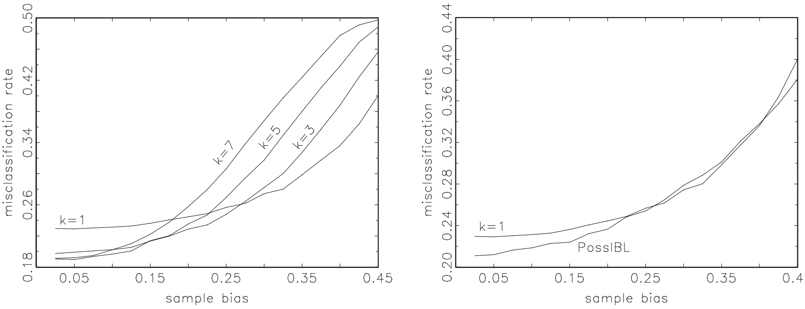


**Fig. 5.6.** Misclassification rates of $k$NN methods (left) and PossIBL (right, in comparison with 1NN).

Apart from $k$NN methods, we have tested PossIBL with $\oplus = \oplus_P$. The similarity measure $\sigma_{\mathcal{S}}$ was defined by the triangle $(x, y) \mapsto \max\{0, 1 - |x - y|/0.8\}$. Interestingly enough, this approach yields the most satisfactory results. For $\alpha$ close to 0 it is almost as good as the $k$NN rules with $k > 1$, and for $\alpha$ close to $1/2$ it equals the 1NN rule. Thus, the combination mode as realized by the probabilistic sum $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ turns out to be reasonable under the conditions of this experiment. As already explained in Section 5.3, this operator produces a kind of saturation effect: It takes frequency information into account, but only to a limited extent (the larger the current support already is, the smaller the absolute increase due to a new observation). Thus, it is indeed in-between the 1NN rule and the $k$NN rules for $k > 1$. Intuitively, this explains our findings in the above experiment, especially that PossIBL is more robust against the sample bias than $k$NN rules for $k > 1$.

Needless to say, what we considered here is only a particular setup in which PossIBL appears to be superior to standard $k$NN with regard to robustness. As robustness is a very multi-faceted aspect, one should not overlook that our results are preliminary and of limited significance.

### 5.5.4 Variation of the aggregation operator

An interesting question concerns the dependence of PossIBL's performance on the specification of the aggregation operator $\oplus$ in (5.13). To get a first idea of this dependence, we have performed the same experiments as described in

Section 5.5.2 above. Now, however, we have tested PossIBL with different t-conorms.

More precisely, we have specified a t-conorm by means of the parameter $\rho$ in (5.18), i.e., we have taken different aggregation operators from the Frank-family of t-conorms. PossIBL was then applied to each data set with different operators $\oplus_\rho$. The results are presented in Appendix E. Each figure shows the average classification performance of PossIBL (over 100 experiments) as a function of the parameter $\rho$. Please note the different scaling of the axes for the five data sets.

Confirming our previous considerations, the results show that in general different t-conorms are optimal for different applications. Still, PossIBL's performance is quite robust toward the variation of the aggregation operator. That is, classification accuracy does not drop off too much when choosing a suboptimal operator.

A very interesting finding is the observation that the parameter $\rho = 0$ and, hence, the maximum operator is optimal if simultaneously the 1NN classifier performs well in comparison with other $k$NN classifiers. If this is not the case as, e.g., for the Balance Scale and the Pima Indians Diabetes data, parameters $\rho > 0$ achieve better results. This finding is not astonishing and can also be grasped intuitively. In fact, it was already mentioned that PossIBL with $\oplus = \oplus_0 = \max$ is closely related to the 1NN classifier, as both methods do fully concentrate on the most relevant information. As opposed to this, aggregation operators $\oplus = \oplus_\rho$ with $\rho > 0$ combine the information from several neighbors in much the same way as do $k$NN classifiers with $k > 1$.

### 5.5.5 Representation of uncertainty

It was already mentioned that an important aspect of PossIBL concerns the representation of uncertainty. The fact that PossIBL can adequately represent the *ignorance* related to a decision problem is easily understood and does not call for empirical validation. To get a first idea of PossIBL's ability to represent *ambiguity* we have derived approximations to two characteristic quantities, again using the experimental setup as described in Section 5.5.1.

Let $D_1$ denote the expected difference (margin) between the possibility degree of the predicted label $r_0^{est}$ and the possibility degree of the second best label, given that the prediction is correct:

$$D_1 \overset{\text{df}}{=} \delta_{s_0}(r_0) - \max_{r \in \mathcal{R}, r \neq r_0} \delta_{s_0}(r). \tag{5.51}$$

Moreover, let $D_0$ denote the expected difference between the possibility degree of the predicted label $r_0^{est}$ and the possibility degree of the actually true label $r_0$, given that $r_0 \neq r_0^{est}$:

$$D_0 \overset{\text{df}}{=} \delta_{s_0}\left(r_{s_0}^{est}\right) - \delta_{s_0}\left(r_{s_0}\right). \tag{5.52}$$

Ideally, $D_0$ is small and $D_1$ is large: Wrong decisions are accompanied by a large degree of uncertainty, as reflected by a comparatively large support of the actually correct label. As opposed to this, correct decisions appear reliable, as reflected by low possibility degrees assigned to all labels $r \neq r_0$.

Table 5.5.5 shows approximations to the expected values $D_0$ and $D_1$, namely averages over $1,000$ experiments. As can be seen, the reliability of a prediction is reflected very well by the possibilistic estimations.

| Dataset | $D_0$ | $D_1$ |
|---|---|---|
| Balance Scale | 0,094 | 0,529 |
| Iris Plant | 0,194 | 0,693 |
| Glass Identification | 0,181 | 0,401 |
| Pima Indians Diabetes | 0,211 | 0,492 |
| Wine Recognition | 0,226 | 0,721 |

**Table 5.6.** Statistics (5.51) and (5.52) for PossIBL.

## 5.6 Calibration of CBI models

The methodological framework introduced in previous sections provides a broad spectrum of techniques for building a CBI model. Needless to say, it would be unrealistic to expect a human expert using these (linguistic) modeling techniques to come up with precise mathematical formalizations of related fuzzy concepts. Instead, a more reasonable approach is to let the expert specify the coarse structure of a model, in our case the fuzzy rules modeling the CBI hypothesis, and to determine the ultimate model in a second step by adapting the expert model to the observed data. This is to some extent comparable, say, to graphical modeling techniques such as Bayesian networks, where the user specifies the structure of the network (i.e., the qualitative part of the model), and the (conditional) probability distributions (i.e., the quantitative part) is learned from data.

In Section 5.4.2, we have already presented a learning scheme for adapting a possibilistic model to the application at hand, albeit for a very particular case (namely PossIBL, our possibilistic variant of IBL). This section is meant to discuss model calibration in more general terms, including the determination of similarity measures and modifier functions. More specifically, we consider the problem of determining modifiers $m_1$ and similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ in a set of rules of the form $m_1 \circ \sigma_{\mathcal{S}} \rightharpoonup \sigma_{\mathcal{R}}$. Each of these rules induces a related possibility distribution (5.7) or, when using aggregation operators other than max and min, the generalized version

$$(s, r) \mapsto \bigoplus_{1 \leq \imath \leq n} m_1(\sigma_{\mathcal{S}}(s, s_\imath)) \otimes \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.53}$$

The overall distribution $\delta_{\mathcal{C}} : \mathcal{S} \times \mathcal{R} \longrightarrow [0,1]$, considered as a lower approximation of the relation $\varphi$ in (5.4), is given by the union (pointwise maximum) of these distributions.

The basic idea is to proceed from similarity measures and modifiers which are specified in the form of parameterized functions. For instance, the modifier associated with the linguistic hedge "very" might be specified by the function $x \mapsto x^{\alpha}$ with $\alpha > 1$. Likewise, the similarity of horsepowers, $\sigma_{hp}$, might be given by the function

$$(x, x') \mapsto \max \left\{ 1 - \frac{|x - x'|}{M}, 0 \right\}, \tag{5.54}$$

where $M$ plays the role of a parameter (cf. Example 5.1). All these parameters can be combined into one vector $\theta$ which determines the CBI model and, hence, has a strong influence on the generalization beyond (via extrapolation of) observed cases. In this sense, it plays a role somewhat similar to, e.g., the smoothing parameter in kernel-based estimation of probability density functions.

In order to determine $\theta$ and, hence, a concrete CBI model from the memory $\mathcal{M}$ of observed cases, a kind of optimization criterion is needed. A reasonable idea is to minimize some distance, such as

$$\int_{\mathcal{C}} \left( \delta_{\mathcal{C}}(c \,|\, \theta) - \delta_{\varphi}(c) \right)^2 \, dc, \tag{5.55}$$

between the estimated distribution $\delta_{\mathcal{C}}(\cdot \,|\, \theta)$ and the (true) $\{0,1\}$-valued distribution $\delta_{\varphi}$ defined by $\delta_{\varphi}(c) = 1 \Leftrightarrow c \in \varphi$.

This is quite comparable with the determination of the *kernel width* or *smoothing parameter* $h$ in kernel-based density estimation, where an underlying density function $\phi$ is estimated by

$$\phi_h : x \mapsto \frac{1}{n} \sum_{\imath=1}^{n} \kappa_h \left( x - x_{\imath} \right) = \frac{1}{n} \sum_{\imath=1}^{n} \kappa \left( \frac{x - x_{\imath}}{h} \right), \tag{5.56}$$

with $\kappa$ being the *kernel function.*[25] The smoothing parameter $h$ has an important effect on the accuracy of the approximation (5.56). It plays a role somewhat similar to the bin-width of histograms. One way of determining this parameter is to minimize the integrated squared error

$$\mathsf{ISE}(h) = \int \left( \phi(x) - \phi_h(x) \right)^2 \, dx \tag{5.57}$$

between the true density $\phi$ and the estimation $\phi_h$.

Unfortunately, (5.57) cannot be derived since the true density $\phi$ is unknown, and the same remark of course also applies to (5.55), where $\pi_{\varphi}(c)$ is not known

[25] Typical examples of $\kappa$ include the Parzen window $u \mapsto \mathbb{I}_{[-1/2,1/2]^m}$ [289] and the normal kernel, the latter being defined as the density of the (multivariate) standard normal distribution.

for all $c \in \mathcal{C}$. A possible way out is to replace the true approximation error by an empirical error, namely the error for the observed cases. This can be done by means of a (leave-one-out) cross validation procedure which, in the case of kernel-based density estimation, approximates the integral by a weighted sum and replaces the density $\phi$ by a further estimation $\widehat{\phi}$ [185]. This leads to the minimization of

$$\sum_{\imath=1}^{n} \left( \widehat{\phi}_h(x_\imath) - \phi_h(x_\imath) \right)^2 , \tag{5.58}$$

where $\widehat{\phi}_h(x_\imath)$ denotes the estimated (cross validation) density for the $\imath$-th observation $x_\imath$. Again, this value is obtained by means of a kernel-based estimation (using $h$ as a smoothing parameter). As opposed to the derivation of $\phi_h(x_\imath)$, however, this estimation leaves the point $x_\imath$ itself out of account, i.e., it uses only the observations $\{x_1, \ldots, x_{\imath-1}, x_{\imath+1}, \ldots, x_n\}$.

The same idea can also be applied to (5.55). In this case, we do not even have to estimate the values $\delta_\varphi(c_\imath)$ since $\delta_\varphi(c_\imath) = 1$ holds true for each observation $c_\imath \in \mathcal{M}$. However, by restricting ourselves to the observed cases, the minimization problem becomes ill-posed. In fact, a trivial solution to the problem of minimizing

$$\sum_{c \in \mathcal{M}} \left( \delta_{\mathcal{C}}(c \,|\, \theta) - \delta_\varphi(c) \right)^2 \tag{5.59}$$

is given by $\delta_{\mathcal{C}}(\cdot \,|\, \theta) \equiv 1$. This simply means to choose the parameter $\theta$ such as to maximize the extrapolation of cases, a hardly convincing result.

In this connection, recall the problem that a possibilistic prediction $\delta_{\mathcal{C}}$ can principally not be "falsified" (cf. Section 5.4.2): The *non-support* of an actually *observed* case can be justified by the fact that no cases have (as yet) been observed which are similar enough. Thus, a small value $\delta_{\mathcal{C}}(c \,|\, \theta)$ is not necessarily a defect of the model, i.e., it does not necessarily indicate a poor choice of the parameter $\theta$. (Predicted possibility degrees are only lower bounds, and low degrees are quite natural if the memory $\mathcal{M}$ does not contain many cases similar to $c$!) Moreover, it is hardly possible to object to the *support* of a yet *unobserved* case since it would require knowledge about the non-existence of that case (which is of course not available). As can be seen, the model based on possibility rules only indicates which cases are (provably) *possible*. It does not, however, point to those cases which appear *impossible*. In other words, the possibilistic model merely expresses the *support* but not the *exclusion* of cases. This contrasts with a probabilistic approach, where an event cannot be supported without (partly) excluding its complement at the same time.

Fortunately, as already pointed out in Sections 5.3.3 and 5.4.2, the (partial) exclusion of cases according to the CBI principle can be realized by means of a complementary type of extrapolation principle induced by a different sort of fuzzy rule, called certainty rule. The latter entails the distribution

$$(s,r) \mapsto \bigotimes_{1 \leq \imath \leq n} (1 - \sigma_{\mathcal{S}}(s,s_\imath)) \oplus \sigma_{\mathcal{R}}(r,r_\imath) \tag{5.60}$$

which actually represents upper bounds and thus defines the counterpart to (5.53). The overall prediction $\pi_{\mathcal{C}}$, associated with a set of rules of that type, is defined by the intersection (pointwise minimum) of the distributions (5.60). As can be seen, a certainty rule reduces the possibility of hypothetical cases which are somehow in conflict with observed cases, in the sense that the inputs are similar but the outcomes are rather different.

Example 5.8. Reconsider Example 5.1 with a case $(100, 15000)$, i.e., a car with horsepower 100 and price \$15,000. In connection with the similar horsepower–similar price hypothesis and the possibility rule model (5.53), this case (partly) *supports* the case $(110, 16000)$ which has a similar horsepower and a similar price. According to the certainty rule model (5.60), it (partly) *excludes* the case $(110, 5000)$ which has a similar horsepower but a rather different price. Observe that the possibility rule model will generally say little about the case $(110, 5000)$, as expressed by a small lower possibility bound. Likewise, the certainty rule model has not much to say about the car $(110, 16000)$ to which it assigns a large upper bound. □

In connection with the determination of optimal similarity measures and modifiers, the two models can complement each other in a reasonable way.[26] As already pointed out in Section 5.3.3, the prediction $\delta_{\mathcal{C}}$ derived from (5.53) and the prediction $\pi_{\mathcal{C}}$ obtained from (5.60) might be *conflicting* in the sense that $\pi_{\mathcal{C}}(c) < \delta_{\mathcal{C}}(c)$ for a case $c$. This can happen if $c$ is supported by some observation $c_1 \in \mathcal{M}$ (according to the possibility rule model) and, at the same time, excluded by another observation $c_2 \in \mathcal{M}$ (according to the certainty rule model). A situation of this kind indicates a defect of the underlying CBI model (the lower possibility bound is larger than the upper bound). It occurs if a case $c$ is similar to both, $c_1$ and $c_2$ (in the sense of the similarity measure $\sigma_{\mathcal{S}}$), and if $c_1$ indicates a result which is quite different (in the sense of $\sigma_{\mathcal{R}}$) from the one suggested by $c_2$. Besides, it should be noticed that a more or less isolated case $c$ does not involve any conflict, since $\delta_{\mathcal{C}}(c)$ and $\pi_{\mathcal{C}}(c)$ will be close to 0 and 1, respectively.

Example 5.9. Suppose, for instance, that we have observed the cars $c_1 = (50, 5000)$, $c_2 = (100, 15000)$, and $c_3 = (75, 7000)$ and that we only distinguish between similar and dissimilar horsepowers resp. prices:

$$\sigma_{\mathcal{S}}(x,y) = \begin{cases} 1 & \text{if} \quad |x-y| \leq \Delta \\ 0 & \text{if} \quad |x-y| > \Delta \end{cases},$$

$$\sigma_{\mathcal{R}}(x,y) = \begin{cases} 1 & \text{if} \quad |x-y| \leq 5000 \\ 0 & \text{if} \quad |x-y| > 5000 \end{cases}.$$

[26] The joint use of lower and upper possibility bounds (derived, respectively, from possibility and certainty rules) has also been advocated in the context of approximate reasoning [376, 393].

For $\varDelta = 30$, $c_1$ qualifies the case $c_3$ as being (completely) possible. However, since $\sigma_{\mathcal{S}}(75, 100) = 1$ as well, $c_3$ is disqualified by $c_2$ at the same time. This suggests to choose a smaller value for $\varDelta$, since otherwise the similar horsepower–similar price rule becomes invalid. More generally, a memory of $n$ cases $\langle s_\imath, r_\imath \rangle$ calls for

$$\varDelta \leq \min_{1 \leq \imath, \jmath \leq n,\, \sigma_{\mathcal{R}}(r_\imath, r_\jmath)=1} |s_\imath - s_\jmath|$$

in order to satisfy this rule. As can be seen, the stronger the variability in the horsepower–price relation is, the more restrictive the similarity between horsepowers has to be defined. In the more general case where similarity measures are not $\{0, 1\}$-valued, a conflict might appear in a less obvious way, and the degree to which the CBI hypothesis is satisfied can vary gradually. □

The above example reveals the following effect: The more similar the cases are made (through the definition of corresponding similarity measures and modifiers), the stronger is the degree of support resp. exclusion induced by a set of observations according to (5.53) resp. (5.60) and, hence, the larger the conflict becomes. Here, we take advantage of this effect in order to define meaningful modifier functions and measures of similarity. In fact, a reasonable optimization criterion is to find a tradeoff between a principle of *appropriate support* (of observed cases) and a *consistency* principle:

- Observed cases should be supported as much as possible by the other cases in the memory (e.g., in connection with a leave-one-out cross-validation).
- The conflict between the support and exclusion of these cases should be as small as possible.

Formally, we define the support attached to a case $c \in \mathcal{M}$ by

$$\mathsf{supp}_\theta(c) \overset{\text{df}}{=} \delta_{\mathcal{C}}(c \,|\, \theta), \tag{5.61}$$

where $\delta_{\mathcal{C}}(\cdot \,|\, \theta)$ is derived from $\mathcal{M} \setminus \{c\}$ according to (5.53) and $m_1, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$ are determined by the parameter vector $\theta$. Moreover, the conflict associated with the case $c$ can be defined as

$$\mathsf{conf}_\theta(c) \overset{\text{df}}{=} \max\{0, \delta_{\mathcal{C}}(c \,|\, \theta) - \pi_{\mathcal{C}}(c \,|\, \theta)\}, \tag{5.62}$$

where $\pi_{\mathcal{C}}(c \,|\, \theta)$ is the distribution obtained from the certainty rule model (5.60). Note that, in the case where possibility is interpreted as an ordinal concept, one might think of replacing the subtraction in (5.62) by a purely qualitative measure of conflict:

$$\mathsf{conf}_\theta(c) = \begin{cases} 1 & \text{if} \quad \pi_{\mathcal{C}}(c \,|\, \theta) < \delta_{\mathcal{C}}(c \,|\, \theta) \\ 0 & \text{if} \quad \pi_{\mathcal{C}}(c \,|\, \theta) \geq \delta_{\mathcal{C}}(c \,|\, \theta) \end{cases} .$$

The derivation of (5.61) and (5.62) for all cases in the memory yields $n$ degrees of support and conflict, respectively. The overall support induced by the parameter $\theta$, $\mathsf{supp}(\theta)$, can then be obtained by aggregating these values:

$$\mathsf{supp}(\theta) = A(\{\mathsf{supp}_\theta(c) \,|\, c \in \mathcal{M}\}) \tag{5.63}$$

with $A$ being an aggregation function. A measure $\mathsf{conf}(\theta)$ of conflict can be defined analogously. Finally, an optimal parameter $\theta$ is derived as a function of the support and the conflict thus defined, e.g., by maximizing

$$\mathsf{supp}(\theta) - \alpha \cdot \mathsf{conf}(\theta) \tag{5.64}$$

for some tradeoff parameter $\alpha \geq 0$ or by maximizing $\mathsf{supp}(\theta)$ under the condition that $\mathsf{conf}(\theta) \leq \alpha$.

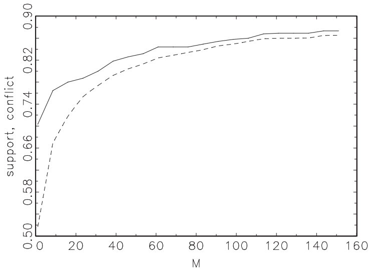


**Fig. 5.7.** Support (solid line) and conflict as a function of the parameter $M$ which defines the similarity measure for the attribute horsepower.

In order to combine the degrees of support (conflict) associated with individual cases, one might use a simple average as an aggregation function $A$ in (5.63). Alternatively, an aggregation which is more in accordance with a qualitative setting is the Sugeno integral

$$\int^{su} \mathsf{supp}_\theta \, d\mu = \sup_{\alpha \geq 0} \min\{\alpha, \mu(F_\alpha)\}, \tag{5.65}$$

where $F_\alpha = \{c \in \mathcal{M} \,|\, \mathsf{supp}_\theta(c) \geq \alpha\}$ for $0 \leq \alpha \leq 1$. The measure $\mu$ in (5.65) can be taken as the counting measure, i.e., $\mu(A) = |A|/|\mathcal{M}|$ for all $A \subseteq \mathcal{M}$.

Example 5.10. Consider as a simple example the choice of the parameter $M$ in (5.54) which defines the similarity measure $\sigma_{hp}$ in connection with the similar horsepower–similar price hypothesis (using the same function with $M = 3000$ for the similarity $\sigma_{\mathcal{R}}$). Fig. 5.7 shows $\mathsf{supp}(M)$ and $\mathsf{conf}(M)$, defined according to (5.61), (5.62), and the aggregation (5.65) as a function of $M$. The choice of $\alpha = 3/4$ in (5.64) suggests $M = 76$ as an optimal parameter and leads to the prediction shown in Fig. 5.8. □

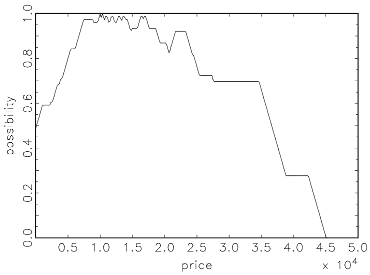


**Fig. 5.8.** Prediction of the price of a car with horsepower 100, where $\sigma_{hp}$ is given by (5.54) with $M = 76$.

Remark 5.11. The calibration method outlined above can be seen as a generalization of related probabilistic approaches. In the latter case, the support and the exclusion of a value always add up to 1. Therefore, a conflict cannot occur, and only the principle of correct support remains relevant. Note that this principle reduces to a principle of *maximal* support in the possibilistic model, as can be gathered from (5.59). In the probabilistic case, the correct support corresponds to the true probability, as expressed by (5.58). □

Let us finally mention that some standard estimation and optimization problems have to be solved in connection with a concrete application. This concerns, for example, the question whether all parameters can be identified by the optimization criterion. Besides, it should be noted that the method of finding an optimal CBI model outlined in this section amounts to solving a nonlinear optimization problem. It might hence be considered critical from the viewpoint of computational complexity, especially since a new parameter has to be derived each time the memory changes. One should realize, therefore, that a parameter estimation is usually not a time-critical problem since it can be solved "off-line." Note that the current optimal parameter can serve as a good initial value when using iterative improvement methods. In fact, a small variation of the memory, such as the adding of a new case, will generally change the optimal parameter but slightly.

## 5.7 Relations to other fields

This section is meant to explore relationships between the possibilistic approach to CBI outlined in previous sections (PoCBI) and some related methods. One can look at PoCBI from different directions. From the viewpoint of statistics and data analysis, it is formally somewhat similar to non-parametric (kernel-based) density estimation. However, as was already discussed in Section 5.3.5, it differs in using possibility theory and similarity instead of probability theory and frequency as major concepts. The use of fuzzy sets and possibility distributions instead of (in addition to) probability distributions is just the characteristic property that PoCBI shares with fuzzy data analysis, the fuzzy set-based counterpart (extension) to classical data analysis. Some relevant aspects of corresponding methods will be discussed in Section 5.7.1.

PoCBI combines rule-based and instance-based reasoning techniques: A memory of cases induces a set of rules and allows CBI to be realized as rule-based reasoning. Besides, Section 5.4.7 has shown that both techniques can be used in a complementary way. The combination of case-based and rule-based reasoning (as well as other hybrid approaches to machine learning) has recently received considerable attention, and it has already led to several interesting approaches [14, 61, 89, 174, 175, 246]. A combined approach is particularly advocated by the complementary merits of the two techniques, namely the suitability for representing general (background) knowledge of a domain in rule induction and specific knowledge in the form of observed cases in CBR. An obvious idea, for instance, is to use a complementary representation in which those cases are stored in the memory which are exceptions to a set of otherwise valid (default) rules. There are, however, other possibilities of combining rule induction and case-based reasoning, some of which have been realized in the Patdex system [18]. PoCBI can be considered from both directions. Since relationships between PoCBI and instance-based learning have already been discussed in Section 5.3.5, this section shall touch on some aspects in connection with more common approaches to fuzzy set-based approximate (rule-based) reasoning.

### 5.7.1 Fuzzy and possibilistic data analysis

The term *fuzzy data analysis* can have different meanings, depending on whether the adjective “fuzzy” refers to the observed *data* itself or to the *methods* used for analyzing the data. That is, a main differentiation must be made between the analysis of somehow uncertain or vague data (e.g., by means of generalized statistical methods [241]) and the use of fuzzy or possibilistic methods for processing data that has been observed precisely (e.g., fuzzy clustering of crisp data [32]). Fuzzy data analysis can also comprise both aspects, of course. It is then concerned with using fuzzy or possibilistic methods for supporting the analysis of vague data [22].

In connection with fuzzy data analysis it is important to distinguish between different types of incomplete knowledge, notably *uncertainty* and *imprecision.* Traditional statistical methods take the first phenomenon into consideration: The generation of data is modeled as a stochastic process, thus leading to random (but still precise) observations. The analysis of fuzzy data does not only consider uncertainty in the generation but also in the observation of data, i.e., it assumes observations to be afflicted with imprecision. In fact, the latter type of uncertainty, which must not be confused with randomness, is often present in practice. Firstly, the observed object itself can be vague in the sense that it might not be possible to identify or demarcate it exactly. Secondly, the measuring instrument or the underlying scale might not allow for identifying the (principally well-defined) object precisely. A standard example is the (linguistic) "value" of a number (which is exact as such) on a scale of linguistic expressions.

Subsequently, we shall briefly discuss some aspects of PoCBI in the context of different approaches to fuzzy data analysis. Qualitative data analysis generally aims at discovering some kind of structure or patterns in the data and, hence, is in line with desriptive statistics, exploratory data analysis, as well as much of current research in the emerging field of data mining and knowledge discovery [183]. Corresponding methods, such as (fuzzy) cluster analysis, mainly focus on single properties of the objects under study and are mainly interested in comparing the data. As in PoCBI, the concept of similarity thus plays a major role in such methods. Besides, PoCBI also helps in getting a more precise idea of the data. To this end, however, it already generalizes beyond the given observations (against the background of further knowledge), whereas qualitative methods consider these observations alone. Seen from this perspective, PoCBI might be considered as an extended form of *exploratory* or *descriptive* data analysis.

While qualitative methods focus on individual properties of an object, *quantitative* analysis is rather concerned with finding (invariant) *relations* between different features, e.g., by estimating (fuzzy) functional relationships (as supervised methods in machine learning).

EXAMPLE 5.12. As a simple example of a quantitative method consider the fitting of a (parameterized) fuzzy set-valued mapping $F_\theta : \mathfrak{R} \longrightarrow \mathfrak{F}(\mathfrak{R})$ to a set of (fuzzy) observations $(x_k, Y_k) \in \mathfrak{R} \times \mathfrak{F}(\mathfrak{R})$ $(1 \leq k \leq n)$. This can be accomplished, e.g., by choosing the (fuzzy) parameter vector $\theta$ such that

$$\sum_{k=1}^{n} \|Y_k - F_\theta(x_k)\|$$

is minimized, where $\| \cdot \|$ is a (metric) distance measure on $\mathfrak{F}(\mathfrak{R})$, the class of fuzzy subsets of $\mathfrak{R}$ [86]. A further possibility is to minimize the spread of $F_\theta$ while somehow covering the data, e.g., while satisfying $Y_k \subseteq F_\theta(x_k)$ for all $1 \leq k \leq n$. The latter type of fuzzy regression analysis amounts to solving a linear programming problem if $F_\theta$ has a certain linear structure. □

Fuzzy methods like the one in Example 5.12 can be interpreted in different ways. Firstly, they can be seen as a generalized approximation (resp. interpolation) method, where scalar observations and functions are replaced by fuzzy set-valued observations and mappings, respectively. Such methods should basically be understood as describing the *given data*, as opposed to inductive statistical methods which draw conclusions about some underlying process which generates the data. For instance, the parameter $\theta$ in Example 5.12 is chosen such that $F_\theta$ fits the data optimally (e.g., in the sense of minimizing the sum of squared errors). It should not be interpreted, however, as an estimation of some true (but unknown) parameter which identifies a data-generating process. Consequently, fuzzy methods of such kind cannot fall back on a related model in order to make predictions. Rather, they have to rely on the same kind of assumptions as CBI, namely that the observations are to some degree representative and that similar outputs are generated by similar inputs [21].[27] It should be observed, however, that the extent of extrapolation (or interpolation) of outputs is principally not bounded, e.g., when fitting a fuzzy mapping to a set of observations and using that mapping for making predictions [87]. Seen from this perspective, corresponding methods seem to lack a solid basis for generalizing beyond observed data.

The use of fuzzy sets for modeling imprecision in the observation of (actually exact) data gives rise to a second interpretation which is related to possibility theory: A fuzzy set $A$ attaches uncertainty to a crisp object (namely its core) and a degree of membership $A(x)$ is considered as the possibility of $x$ being the true (only incorrectly observed) object. This interpretation has motivated the introduction of *possibilistic variables* as a counterpart to random variables. The related idea of a *possibilistic* generation of data leads to parameter estimation methods which parallel the maximum likelihood estimator in statistics (by using the minimum operator instead of the product) [22]. Corresponding methods thus fall into line with model-based approaches in mathematical statistics. Since each observation induces a possibility distribution $\pi = A$, this type of modeling is closely related to PoCBI. Still, the underlying semantics is very different. In the first case, *indistinguishability* is taken as a necessary evil, and $A(x)$ quantifies the possibility that the real object, $x_0$, is *actually given* by $x$. In the second case, *similarity* is exploited as a useful concept for pointing to the existence of other objects, and $\pi(x)$ is considered as the plausibility of encountering $x$ (while knowing the current object $x_0$). As a further difference let us mention that the ensemble of fuzzy observations (the possibilistic data set) marks the *input* in possibilistic data analysis. It is further processed by means of generalized methods, such as possibilistic linear regression [367] or possibilistic cluster analysis [191]. In PoCBI, the union of possibility distributions principally corresponds to the output, whereas the input is given in the form of precise cases.

A third interpretation of fuzzy methods establishes a close connection between fuzzy sets (fuzzy data) and probability theory and makes use of concepts such

[27] Indeed, this assumption is implicitely made when fitting a *continuous* (fuzzy) mapping.

as like probabilistic sets [189], fuzzy random variables [241] or random fuzzy sets [303]. This approach calls for generalizations of classical statistical methods. It also leads to possibilistic reasoning methods which can be seen as a kind of approximate probabilistic inference. Let us mention the learning of possibilistic networks from data which is based on a probabilistic interpretation of possibility degrees (in terms of random sets) as an example [42, 43]. Possibilistic networks emerge from probabilistic networks (including Bayesian networks [292] and Markov networks [245]) by using possibility distributions instead of probability measures. This allows one to take uncertainty as well as imprecision into account [41]. Apart from the probabilistic semantics, they can hence be seen as the possibilistic counterpart to probabilistic networks in much the same way as PoCBI can be considered as the possibilistic counterpart to kernel-based density estimation.

Graphical modeling by means of network structures is an example of a model-based approach which is capable of combining knowledge and data in various ways, a property which is often emphasized as a major benefit [187]. Typically, an expert specifies the structure of a network, i.e., the qualitative part, while the associated (conditional) probability or possibility distributions are learned from data. Compared with the use of rules (which define the qualitative part of the model in PoCBI), knowledge hence appears in the form of (in)dependence relations between variables represented by means of a directed (acyclic) graph. Besides, the (conditional) probabilities or possibilities, i.e., the quantitative part of a network, correspond to the similarity measures and modifier functions in PoCBI, which can be adapted to observed data by means of corresponding learning method (cf. Section 5.4.2).

In summary, PoCBI has characteristics in common with both, qualitative and quantitative data analysis. It is close to qualitative approaches in making use of similarity as a basic concept and in supporting the description of data. Still, it is also concerned with generalizing and making predictions, a property it shares with possibilistic approximation or parameter estimation. As opposed to PoCBI, however, such methods are mostly model-based. Besides, the meaning of a possibility distribution in PoCBI greatly differs from the interpretation in the methods outlined in this section, the latter using such distributions for modeling uncertain or vague data, parameters or predictions.

### 5.7.2 Fuzzy set-based approximate reasoning

Fuzzy rule-based modeling and related approximate reasoning techniques are among the most popular applications of fuzzy set theory. Fuzzy rules have been used extensively for the linguistic modeling of functional relationships. The main idea of fuzzy control, for instance, is to simulate a human expert by constructing a control function from a set of linguistically specified *if-then* rules. In this context, a rule “if $X$ is $A$ then $Y$ is $B$” represents (vague) partial knowledge about the graph of an underlying (control) function and is usually not considered as a logical implication. Rather, it defines an (ordered) pair of (fuzzy) data $(A, B)$ and should be understood in the sense of a possibility-qualifying rule. The union of fuzzy relations $A \times B$ associated with a number of rules defines a *fuzzy graph* [418]. It is thought of as a vague approximation of the underlying (control) function in much the same way as $\delta_{s_0}$ is interpreted as a (lower) approximation of the relation $\varphi$ of cases.

Seen from this perspective, PoCBI is close to the interpretation of fuzzy rules originally outlined by Zadeh [414] and put into practice by Mamdani [259, 258]. Still, a major difference deserves mentioning: A human expert specifying points of the graph of a function is assumed to have knowledge about *absolute* values of that function. By providing similarity-based rules in PoCBI, he rather gives a description of how these values vary when changing the argument of the function. For example, an expert might know very little about prices of cars of a certain manufacturer. Still, his (case-based) experience might tell him that (at least in general) cars with similar horsepower and similar engine-size have similar prices. Then, learning about the price of one (typical) car of a certain manufacturer, he will also have an idea of the price of a similar car (produced by the same manufacturer).

Mathematically speaking, PoCBI assumes that a human expert can somehow specify, not a function itself, but the variation or derivative of the function. This knowledge can then be used for extrapolating observed data in the form of concrete values. By instantiating observed cases, PoCBI thus transforms a set of similarity-based rules into a (larger) set of ordinary fuzzy rules. In other words, an ordinary rule base is derived from a set of similarity-based rules in connection with a set of observations. Needless to say, this type of case-based derivation of a rule base might be interesting not only for CBR itself but also for other domains. In fuzzy control, for instance, it might reasonably complement other techniques for learning fuzzy rules (e.g. [2, 386]). In this sense, PoCBI can be seen from two perspectives. Firstly, as a method which makes use of fuzzy set-based modeling techniques in order to specify a CBR model, i.e., as an application of fuzzy set (possibility) theory in case-based reasoning. Secondly, as a method which allows one to transform case-based information into a fuzzy rule base, i.e., as an application of CBR techniques in (rule-based) approximate reasoning.

Of course, if the expert is also able to specify some values of a function it seems reasonable to combine PoCBI and the approach to approximate reasoning used in fuzzy control, an idea which has already been discussed in Section 5.4.7. Besides, it should be mentioned that a rule base thus obtained can be “tuned” in different ways. For instance, in order to reduce the size of the case base it will often be reasonable to merge several rules which originate from similar cases, i.e., to derive one general rule from a number of more specific rules (see, e.g., [406] and Section 5.4.3).

## 5.8 Summary and remarks

### Summary

- In this chapter, we have outlined a possibilistic approach to case-based inference. The basic principle of this approach, referred to as PoCBI, is a kind of similarity-guided, possibilistic extrapolation of observed cases. According to this principle, which relies on the CBI hypothesis and which has been formalized within the framework of fuzzy rules, an already encountered case is taken as evidence for the existence of similar cases. This evidence is expressed in terms of degrees of possibility assigned to hypothetical cases and thus defines a possibilistic approximation of an underlying (but only partially observed) set of potential cases.
- A distinctive feature of PoCBI is the ability to combine knowledge and data in a flexible way. Even though it can be considered as a case-based method in the first place, (expert) knowledge still plays an essential role. Firstly, such knowledge is used for controlling the "possibilistic extrapolation" of sample cases, i.e., the local generalization beyond observed examples. Secondly, general background knowledge can supplement case-based information when it comes to making predictions. A prediction in the form of a possibility distribution may thus result from the combination of several ingredients, namely the observed cases, the (heuristic) "CBR knowledge" which dictates how to extrapolate the data, and background knowledge which supplements or modifies the extrapolation.
- One of the basic ideas of our approach is that of exploiting the merits of linguistic modeling techniques in the context of CBR. It does not mean, however, that a human expert is expected to come up with an optimal model from the start. Rather, it might be sufficient if he specifies a broad structure in a first step, including, e.g., the selection and combination of important attributes which appear together in a rule. A corresponding rule base can then be calibrated afterwards by means of the adaptation technique proposed in Section 5.6.
- From a learning point of view, the possibilistic approach has much in common with non-parametric statistical inference (kernel-based density estimation) and instance-based learning. In fact, the application of possibility theory allows for realizing a graded version of the similarity-based extrapolation principle underlying IBL which appears to be very natural and intuitively appealing. We have presented a detailed comparison of the possibilistic extrapolation principle and the commonly used approach which can be endowed with a probabilistic basis. Even though the two methods are based on quite different semantics, the possibilistic variant (PossIBL) can formally be seen as an extension of the probabilistic approach. Indeed, it has been shown that the former – at least in its general form – can mimic the latter. Apart from that, the possibilistic approach seems to have some advantages:

From a *knowledge representation* point of view, a possibilistic (instance-based) prediction is more expressive than a probabilistic one. Especially, the former is able to represent the *absolute* amount of evidential support as well as partial ignorance, a point which seems to be of major importance in IBL. Furthermore, the interpretation of aggregated degrees of individual support in terms of (guaranteed) possibility (degrees of confirmation) is generally less critical than the interpretation in terms of degrees of probability.

Regarding the *applicability*, the possibilistic approach is more robust and may thus extend the range of applications. Particularly, it makes no statistical assumptions about the generation of data and less mathematical assumptions about the structure of the underlying instance space. In fact, it was shown that PossIBL performs at least as well as standard NN techniques for typical (real-word) data sets. Beyond that, however, it can also be applied to data that violates certain statistical assumptions. Also worth mentioning is that the max–min version of PossIBL can even be applied within a purely ordinal setting.

Finally, the possibilistic method is more *flexible* and supports several *extensions* of IBL. This includes the adaptation of aggregation modes in the combination of individual degrees of support, the coherent handling of incomplete information, and the graded discounting of atypical cases. Moreover, it allows one to complement the similarity-based extrapolation principle by other inference procedures.

- PoCBI is also related to possibilistic data analysis. In this regard, it was found that it combines aspects of qualitative (descriptive, exploratory) and quantitative (inductive) methods and that it can be seen as a kind of extended exploratory data analysis. The comparison between PoCBI and fuzzy set-based approximate reasoning has shown that PoCBI applies fuzzy rules at a higher level. In connection with observed data, a set of such rules induces or, say, instantiates an “ordinary” (fuzzy) rule base. Thus, case-based and rule-based reasoning techniques can complement each other in a reasonable way.

## Remarks

- The type of possibilistic prediction realized by PoCBI can be used in various ways, e.g., as in this chapter for classification or function approximation. Besides, it can be embedded into more complex reasoning procedures. In the context of case-based reasoning, for example, PoCBI can support the overall process of problem solving by bringing a set of potential solutions into focus: By providing estimations $\delta_{s_0}(r)$ of the possibility that $r$ is the solution (= outcome) of the new problem (= input) $s_0$, or that $r$ can at least be adapted in a suitable way, PoCBI allows one to focus on the most promising candidates and, hence, to improve the efficiency of case-based problem solving. Likewise, a prediction in the form of a possibility distribution can provide useful information in the context of decision making (cf. Chapter 7).

- A case is often characterized by a set of attributes, and a similarity relation is given for each of these attributes (cf. Section 2.3.3 and Example 5.1). In this connection, it deserves mentioning that the derivation of a global similarity by means of an aggregation of individual similarity relations presupposes the individual measures to be *commensurate*: Given two measures $\sigma_1$ and $\sigma_2$ ranging on (numeric) scales $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively, the objects $x$ and $y$ are as similar as $u$ and $v$ iff the equality $\sigma_1(x, y) = \sigma_2(u, v)$ holds. This remark is particularly important in connection with ordinal scales (which might even have different cardinalities). At a formal level, commensurability can also be achieved by mapping similarity degrees from different (heterogeneous) scales into one common scale $\mathcal{L}$ before aggregation takes place.
- We have stressed the aspect that a possibilistic CBI model is essentially derived from the knowledge of an expert, and that data is only used for calibrating the model. Of course, other approaches which partly rely on user advice in model building exist as well, but often the user plays a less significant role or intervenes in a more indirect way. In the memory-based reasoning methodology presented in [217], for instance, the user can specify causal dependencies between variables by (partially) determining the structure of a probabilistic network. This network (eventually in a corrected form) is then used for deriving a similarity-measure which in turn controls the retrieval of cases (and, hence, the labeling of new cases in a classification task).
- Note that the possibilistic approximation of the relation $\varphi$ in (5.4) will in general not converge toward (the $\{0, 1\}$-valued possibility distribution associated with) $\varphi$ with an increasing sample size. Rather, some hypothetical cases similar to observed cases will always be supported with a positive degree of similarity even though they do actually not exist. This problem could be alleviated by controlling the extent of extrapolation as a function of the sample size.[28] This is comparable to a corresponding adaptation of the smoothing parameter in kernel-based density estimation. Notice, however, that an adaptation of this kind is already realized by the calibration of a CBI model (cf. Section 5.6), albeit in a more implicit way. Besides, it should be mentioned that an asymptotic influence of similarity might indeed be reasonable. It makes sense, e.g., if the sample is not representative and some cases are not accessible to observation [281].
- The generalization of the $k$Nearest Neighbor algorithm which has been proposed in [84] is also closely related to the possibilistic approach of this chapter. As already explained in Section 4.9, this approach specifies the unknown class $c_0$ of a new pattern $x_0$ in terms of a belief function. This belief function is

[28] The opinion that the influence of similarity should decrease if the sample size increases was already held by Carnap in connection with the inductive logic-based modeling of analogical reasoning [60].

obtained by combining the individual belief functions induced by the neighbors of $x_0$, where the $\imath$-th neighbor $x_\imath$ specifies $c_0$ by means of a mass distribution $\mathsf{m}_\imath$ such that

$$\mathsf{m}_\imath(\{c_\imath\}) = \alpha_\imath, \quad \mathsf{m}_\imath(C) = 1 - \alpha_\imath. \tag{5.66}$$

Note that the belief structure (5.66) is consonant, which means that it can also be expressed in terms of a possibility distribution.

The main differences between [84] and PoCBI are as follows: Firstly, the combination of individual pieces of evidence is realized in different ways, namely by means of a $\oplus$-aggregation in PoCBI and by means of Dempster's rule in [84]. Note that the latter assumes the pieces of evidence to be distinct [349] which, as argued in Chapter 4, might not always be true in the context of classification.

Secondly, as in IBL, the method in [84] does not consider a similarity structure over the set of outcomes (classes). In fact, an instance only supports the class to which it belongs. As opposed to this, a case also supports *similar* outcomes in PoCBI.

Thirdly, by focusing on classification as a performance task, the method in [84] has been developed with a specific application in mind and can be seen as a purely data-driven approach. As has been seen in previous sections, PoCBI supports the combination of data and domain-specific (expert) knowledge in the more general context of case-based reasoning. This becomes possible through the close connection between possibility theory and the theory of fuzzy sets. In particular, this connection allows one to adapt a possibilistic CBI model by means of fuzzy set-based (linguistic) modeling techniques.

– When comparing the extrapolation principle of the possibilistic and the probabilistic NN principle (Section 5.3.5) we have emphasized the difference between absolute and relative support of a case. A similar distinction has also been made in the context of clustering. In fuzzy clustering, a point is not assigned to one class in an unequivocal way; rather, it may have a positive degree of membership in several classes. Still, in the classical approach the membership degrees are forced to sum to 1 [32]. Consequently, these membership degrees must be interpreted as *relative* numbers. This constraint (which has a probabilistic flavor) is relaxed in *possibilistic* clustering [240], where a membership degree does indeed reflect the (absolute) compatibility of a point with the prototype of a cluster.

– In the qualitative (max-min) version of PoCBI, the evidential support of a hypothetical case $c$ basically corresponds to the maximal similarity between $c$ and an observed case. Interestingly enough, the same value also plays an important role in a probabilistic model of analogical induction proposed in [281]. This value, which corresponds to the possibility degree (5.7) in our approach, is called *analogy factor*.[29] In [281], however, this factor is not directly considered as a measure of evidence. Rather, it is used for modeling the influence of experience from similar situations when it comes to *updating* a degree of probability (of occurrence) associated with $c$.

[29] More precisely, it is qualified as an *existential* analogy factor. An *enumerative* factor which depends on the similarity of $c$, not only to the nearest neighbor, but to *all* observed cases is considered as an alternative.

# 6. Fuzzy Set-Based Modeling of Case-Based Inference II

In Chapter 5, it has already been shown that fuzzy rules can be modeled formally as possibility distributions constrained in terms of a combination of the membership functions which define, respectively, their antecedent and consequent part. This way, they relate the concepts of similarity and uncertainty, which is the main reason for their convenience as formal models of the CBI hypothesis. Work on fuzzy *if–then* rules has mainly concentrated on algebraic properties of (generalized) logical operators. However, going into the semantics of such rules, it turns out that different interpretations lead to different types of fuzzy rules, which can be associated with corresponding classes of implication operators [117].

The logical operator used for modeling the type of fuzzy rule that we have focused on in Chapter 5, a so-called possibility rule, is a conjunction (t-norm) rather than an implication. In fact, a possibility rule is considered, not as a logical implication in the strict sense, but rather as an example-oriented rule which encodes and extrapolates information derived from observations. In this context, a fuzzy rule "if $X$ is $A$ then $Y$ is $B$" defines a case in the form of an ordered pair of data $(A, B)$ which suggests the feasibility of further (similar) cases (or, more precisely, guarantees a certain degree of possibility of such cases).

As already pointed out, however, an alternative, *implication-based* type of fuzzy rule can be very useful in the context of CBI, both from a knowledge representation (Section 5.3.3) and a learning point of view (Section 5.6). In this chapter, we shall consider implication-based fuzzy rules in more detail. As will be seen, formalizing the CBI hypothesis in terms of implication-based rules involves a completely different approach to knowledge representation and inference. In fact, the use of implication-based fuzzy rules leads to a *constraint-based* approach which can be seen as a generalization of the constraint-based modeling of CBI in Chapter 3. That is, each rule associated with an observed case $\langle s_1, r_1 \rangle$ serves as a constraint: Given a new input $s_0$ similar to $s_1$, it rules out those outcomes which are not sufficiently similar to $r_1$. This way, an observation restricts the set of possible outputs resp. decreases the possibility of certain outcomes. Loosely speaking, a constraint-based (implication-based) fuzzy rule *excludes* outcomes which are *dissimilar* (while not saying anything about the similar ones), whereas an example-oriented (conjunction-based) rule *supports* outcomes which are *similar* (while reserving judgement concerning the ones which are dissimilar). The difference between the two approaches, which exactly corresponds to the distinction between certainty and plausibility resp. upper and lower possibility in Section 5.1, becomes also apparent from the way in which evidence from multiple cases is combined. In connection with implication-based rules, this evidence is aggregated by means of an intersection (resp. the application of a t-norm) which is a natural approach to combining constraints. As opposed to this, the disjunctive aggregation (resp. the application of a t-conorm) in the case of possibility rules corresponds to a data accumulation process.

The remaining part of the chapter is organized as follows: In Section 6.1 and Section 6.2, two basic models which make use of two types of implication-based fuzzy rules, namely *gradual rules* and *certainty rules*, are introduced. Section 6.3 considers case-based inference in the context of information fusion and provides a probabilistic interpretation which relates the gradual rule and the certainty rule model. The rating of cases based on the information they provide and the related idea of "exceptionality" of cases is considered in Section 6.4. Section 6.5 generalizes the previously introduced models by applying the CBI hypothesis in a locally resticted way.

Before going on, let us make a note on notation. As in Chapter 5, we shall denote by $\varphi \subseteq \mathcal{S} \times \mathcal{R}$ the set of potential observations, i.e., a case is always an element of the relation $\varphi$. Alternatively, we shall look at $\varphi$ as a set-valued mapping $\varphi : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$, i.e., we denote by $\varphi(s)$ the set $\varphi \cap (\{s\} \times \mathcal{R})$ of possible outcomes of the input $s$. We shall further abuse this notation and write $r = \varphi(s)$ instead of $(s, r) \in \varphi$ or $\{r\} = \varphi(s)$ if $\varphi$ is an ordinary function. Again, we assume data to be given in the form of a (finite) memory

$$\mathcal{M} = \big\{ \langle s_1, r_1 \rangle, \langle s_2, r_2 \rangle, \ldots, \langle s_n, r_n \rangle \big\}$$

of precedent cases. Let $\mathcal{M}^*$ denote the class of all finite memories $\mathcal{M} \subset \varphi$.

Finally, we restrict ourselves in this chapter to the qualitative version of possibility theory and, hence, to the operators min and max as t-norm and t-conorm, respectively. Thus, we assume that possibility (and hence similarity) is measured on an ordinal scale $\mathcal{L}$. (Though an exception is made in Section 6.3, where a possibilistic prediction is endowed with a probabilistic semantics.) We note, however, that all results can be transferred to the quantitative case in a more or less straightforward way.

## 6.1 Gradual inference rules

### 6.1.1 The basic model

Gradual rules [119] depict relations between variables $X$ and $Y$ which correspond to propositions of the form "the more $X$ is $A$, the more $Y$ is $B$," where $A$ and $B$ are fuzzy sets modeling certain symbolic labels. This can also be stated as "the larger the degree of membership of $X$ in the fuzzy set $A$, the larger the degree of membership of $Y$ in $B$" or, even more precisely, as "the larger the degree of membership of $X$ in the fuzzy set $A$, the larger the guaranteed lower bound to the degree of membership of $Y$ in $B$." The intended semantics of such a rule can be expressed in terms of membership degrees by

$$A(X) \leq B(Y), \tag{6.1}$$

which is equivalent to the collection of constraints

$$\forall\, 0 < \alpha \leq 1 \;:\; X \in A_\alpha \Rightarrow Y \in B_\alpha,$$

where $A_\alpha = \{x \,|\, A(x) \geq \alpha\}$ denotes the $\alpha$-cut of the fuzzy set $A$ [119].

The constraint (6.1) induces a $\{0,1\}$-valued (conditional) possibility distribution $\pi_{Y|X}$, where $\pi_{Y|X}(y \,|\, x)$ denotes the possibility of $Y = y$ given that $X = x$:

$$\forall\, x \in D_X \,\forall\, y \in D_Y \;:\; \pi_{Y|X}(y \,|\, x) = A(x) \overset{\text{rg}}{\rightsquigarrow} B(y), \tag{6.2}$$

where $\overset{\text{rg}}{\rightsquigarrow}$ is the Rescher-Gaines implication ($\alpha \overset{\text{rg}}{\rightsquigarrow} \beta = 1$ if $\alpha \leq \beta$ and 0 otherwise) and $D_X$ and $D_Y$ are the domains of $X$ and $Y$, respectively.

More generally, fuzzy gradual rules can be classified as *truth-qualifying rules*, the semantics of which are adequately modeled by means of so-called R(esiduated)-implications. An R-implication is derived from a t-norm $\otimes$ through residuation [118]:

$$\forall\, \alpha, \beta \in [0,1] \;:\; \alpha \rightsquigarrow \beta \overset{\text{df}}{=} \sup\{\, \gamma \,|\, \alpha \otimes \gamma \leq \beta \,\}. \tag{6.3}$$

An example is the implication operator $\rightsquigarrow$ defined as

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 \text{ if } \alpha \leq \beta \\ \beta \text{ if } \alpha > \beta \end{cases} .$$

Using this implication, the possibility of $Y = y$ is not restricted to the values 0 and 1 but may take any value in the interval $[0,1]$. Nevertheless, subsequently we will adhere to the model (6.2) which is referred to as a *pure gradual rule* in [46].

Within the context of our CBI framework, a gradual rule reads "the more similar two inputs are, the more similar are the associated outcomes" or, more precisely, "the more the similarity of inputs is in $F$, the more the similarity of outcomes is in $G$," with $F$ and $G$ being fuzzy sets of "large similarity degrees" ($F$ and $G$ are non-decreasing $\mathcal{L} \longrightarrow \mathcal{L}$ functions). In connection with (6.1) and an observed case $\langle s_1, r_1 \rangle$, this rule (completely) excludes the existence of other (hypothetical) cases $\langle s, r \rangle$ which would violate

$$F(\sigma_{\mathcal{S}}(s, s_1)) \leq G(\sigma_{\mathcal{R}}(r, r_1)). \tag{6.4}$$

Thus, given a new input $s_0$ and assuming $F = G = \text{id}$, (6.4) becomes

$$\forall \langle s, r \rangle \in \varphi \, : \, \sigma_{\mathcal{S}}(s, s_1) \leq \sigma_{\mathcal{R}}(r, r_1) \tag{6.5}$$

and, hence, leads to the restriction

$$r_0 \in \big\{ r \in \mathcal{R} \, | \, \sigma_{\mathcal{S}}(s_0, s_1) \leq \sigma_{\mathcal{R}}(r, r_1) \big\} \tag{6.6}$$

for the output $r_0$ associated with $s_0$. Since corresponding constraints are obtained for all cases of a memory $\mathcal{M}$, we finally derive the following prediction [99, 101]:

$$r_0 \in \widehat{\varphi}_{\mathcal{M}}(s_0) \overset{\text{df}}{=} \bigcap_{1 \leq \imath \leq n} \big\{ r \in \mathcal{R} \, | \, \sigma_{\mathcal{S}}(s_0, s_\imath) \leq \sigma_{\mathcal{R}}(r, r_\imath) \big\} \, . \tag{6.7}$$

Clearly, the extent to which the CBI hypothesis holds true depends on the respective application. Consequently, the formalization of this principle by means of the constraint (6.1) might be too strong, at least in connection with the underlying similarity relations $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$. That is, cases $\langle s, r \rangle, \langle s', r' \rangle$ might exist such that $\sigma_{\mathcal{S}}(s, s') > \sigma_{\mathcal{R}}(r, r')$, i.e., although the inputs are similar to a certain degree, the same does not hold for the associated outputs. This, however, contradicts (6.4). Thus, calling a prediction $\widehat{\varphi}_{\mathcal{M}}(s_0)$ *correct* (with respect to the case $\langle s_0, r_0 \rangle$) if $r_0 \in \widehat{\varphi}_{\mathcal{M}}(s_0)$, the (general) correctness of the inference scheme (6.7) is not guaranteed in the sense that it might yield an incorrect prediction:

$$\exists \, \mathcal{M} \in \mathcal{M}^* \, \exists \, \langle s_0, r_0 \rangle \in \varphi \, : \, r_0 \not\in \widehat{\varphi}_{\mathcal{M}}(s_0).$$

That is, there are a memory $\mathcal{M}$ and a case $\langle s_0, r_0 \rangle$ such that the set-valued prediction derived from $\mathcal{M}$ does not cover $r_0$. Note that the complete class $\varphi$ of cases would have to be known in order to guarantee the correctness of (6.7) in the above sense. Needless to say, this condition is usually not satisfied.

### 6.1.2 Modification of gradual rules

Again, more flexibility can be introduced in the basic model (6.1) by means of a modifier, i.e., a non-decreasing function $m : \mathcal{L} \longrightarrow \mathcal{L}$. This leads to

$$\forall \langle s, r \rangle \in \varphi \, : \, m(\sigma_{\mathcal{S}}(s, s_1)) \leq \sigma_{\mathcal{R}}(r, r_1) \tag{6.8}$$

instead of (6.5). Moreover, (6.7) becomes

$$r_0 \in \widehat{\varphi}_{m,\mathcal{M}}(s_0) \overset{\text{df}}{=} \bigcap_{1 \leq \imath \leq n} \big\{ r \in \mathcal{R} \, | \, m(\sigma_{\mathcal{S}}(s_0, s_\imath)) \leq \sigma_{\mathcal{R}}(r, r_\imath) \big\} \, . \tag{6.9}$$

The application of the modifier $m$ can be seen as "calibrating" the similarity scales underlying the set of inputs and the set of outputs such that (6.1) is always satisfied. As an extreme example of (6.8) consider the case where $m \equiv 0$, expressing the fact that the CBI hypothesis does not apply at all. In other words, the similarity of inputs (in the sense of $\sigma_{\mathcal{S}}$) does not justify any conclusions about the similarity of outcomes (in the sense of $\sigma_{\mathcal{R}}$). Observe, however, that $m$ can as well be utilized in order to strengthen (6.1). We might take, for instance, $m \equiv 1$ if all outcomes are always perfectly similar according to $\sigma_{\mathcal{R}}$! This type of modification of a gradual rule can be interpreted in the same way as the modification of a possibility rule (cf. Section 5.4.4).

We call a modifier *admissible* if it guarantees the correctness of the inference scheme (6.9), i.e.

$$\forall \mathcal{M} \in \mathcal{M}^* \, \forall \langle s_0, r_0 \rangle \in \varphi \, : \, r_0 \in \widehat{\varphi}_{m,\mathcal{M}}(s_0). \tag{6.10}$$

The modifier $m$ defined by

$$m(x) = \sup \big\{ h(x') \,|\, x' \in D_{\mathcal{S}}, x' \leq x \big\} \tag{6.11}$$

for all $x \in D_{\mathcal{S}}$, where

$$h(x) = \inf_{\langle s,r \rangle, \langle s',r' \rangle \in \varphi : \sigma_{\mathcal{S}}(s,s')=x} \sigma_{\mathcal{R}}(r,r'),$$

is admissible. Moreover, it is maximally restrictive in the sense that

$$\forall \mathcal{M} \in \mathcal{M}^* \, \forall s_0 \in \mathcal{S} \, : \, \widehat{\varphi}_{m,\mathcal{M}}(s_0) \subseteq \widehat{\varphi}_{m',\mathcal{M}}(s_0)$$

holds true for each admissible (and non-decreasing) $m' : D_{\mathcal{S}} \longrightarrow \mathcal{L}$.[1] Taking the upper bound in (6.11) only guarantees that $m$ is non-decreasing. In fact, (6.10) remains valid when replacing $m$ by $h$, which obviously corresponds to the *similarity profile* as introduced in Section 3.1.[2] In other words, a modifier $m$ defines a *strict similarity hypothesis* (see page 61) and thus obeys the "the more... the more..." assumption underlying the concept of a gradual rule: The modification by means of a non-decreasing function corresponds to the "stretching" and "squeezing" of the similarity scale underlying $\sigma_{\mathcal{S}}$. When interpreting $m \circ \sigma_{\mathcal{S}}$ as a new (adapted) similarity measure, $m \circ \sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ are still *coherent* in the sense that

$$\sigma_{\mathcal{S}}(s_1, s_2) \leq \sigma_{\mathcal{S}}(s_3, s_4) \;\Rightarrow\; m(\sigma_{\mathcal{S}}(s_1, s_2)) \leq m(\sigma_{\mathcal{S}}(s_3, s_4)) \tag{6.12}$$

for all $s_1, s_2, s_3, s_4 \in \mathcal{S}$. As opposed to this, a non-increasing function $h$ also puts the similarity degrees $x \in D_{\mathcal{S}}$ in a different order and, hence, violates (6.12).

Loosely speaking, (6.11) can be seen as a solution to the (optimization) problem of finding a modifier maximally restrictive among all the admissible ones. Estimating (6.11) from observed data (in the form of the memory $\mathcal{M}$) can be considered as a problem of *case-based learning*. Of course, a corresponding estimation will generally not allow for verifying the admissibility of a modifier in the sense of (6.10). In fact, (6.10) can be checked only for the *observed* cases, which means

[1] Here, we assume that $m'(x) \in \mathcal{L}$ for all $x \in D_{\mathcal{S}}$. More generally, a modifier is a $D_{\mathcal{S}} \longrightarrow [0,1]$ mapping.

[2] Recall, however, that $\varphi$ as defined here is not necessarily a functional relation.

that the requirement of (global) admissibility has to be weakened. An obvious idea is to look for a maximally restrictive modifier $m$ which is admissible, not necessarily for the complete relation $\varphi$, but at least for the memory $\mathcal{M}$. That is,

$$\forall \langle s, r \rangle \in \mathcal{M} \, : \, r \in \widehat{\varphi}_{m,\mathcal{M}}(s). \tag{6.13}$$

In addition to (6.13), it might appear natural to require

$$\forall s \in \mathcal{S} \, : \, \widehat{\varphi}_{m,\mathcal{M}}(s) \neq \emptyset. \tag{6.14}$$

That is, for each input $s$ which might be encountered, the inference scheme (6.9) yields a non-empty (even if perhaps incorrect) prediction [100]. Needless to say, the additional requirement (6.14) makes the learning of a modifier more complex.[3] Note that the problem of learning the maximally restrictive modifier (6.13) can be approached by the algorithm proposed in Section 3.4 (cf. Remark 3.31).

Observe that $F = G = \text{id}$ can be assumed for the fuzzy sets $F$ and $G$ in (6.4) without loss of generality (as long as $G$ is strictly increasing). This becomes obvious from the constraint (6.8). Namely, $m(F(\sigma_{\mathcal{S}}(s, s')) \leq G(\sigma_{\mathcal{R}}(r, r'))$ is equivalent to $m'(\sigma_{\mathcal{S}}(s, s')) \leq \sigma_{\mathcal{R}}(r, r')$ with $m' = G^{-1} \circ m \circ F$.

Even though the approach (6.8) allows for the adaptation of the formal CBI model based on a gradual rule, this model remains rather restrictive. In fact, the above discussion has shown that the gradual rule model is closely related to the constraint-based approach of Chapter 3.[4] Consequently, it might lead to imprecise predictions for exactly the same reasons. Consider the following example, to which we shall return occasionally in subsequent sections.

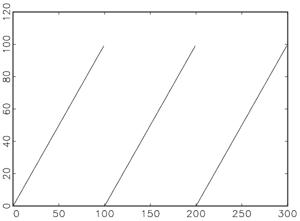


**Fig. 6.1.** Graph of the function $a \mapsto a \bmod 100$.

[3] Verifying (6.14) is closely related to testing the *coherence* of a set of gradual rules [133].

[4] The approaches basically differ in the sense that the latter does not only allow for strict similarity hypotheses.

Example 6.1. Let a CBI setup be defined as follows:

$$\mathcal{S} = \mathcal{R} = \mathfrak{N}_0, \quad D_{\mathcal{S}} = D_{\mathcal{R}} = \{0,1\},$$
$$\sigma_{\mathcal{S}}(a,b) = \sigma_{\mathcal{R}}(a,b) = 1 \Leftrightarrow |a-b| \leq 10,$$
$$\varphi : \mathcal{S} \longrightarrow \mathcal{R}\,,\ a \mapsto a \bmod M.$$

Thus, inputs and outputs correspond to natural numbers, and two inputs (outputs) are either completely similar or not similar at all. According to the definition of $\varphi$,

$$\varphi(a) = q \Leftrightarrow q \in \{0,1,\ldots,M-1\} \\ \wedge\ \exists\, p \in \mathfrak{N}_0\ :\ a = pM + q.$$

Assuming $M$ to be a rather large integer, we can hence say that $\varphi(s)$ and $\varphi(s')$ are "almost surely" similar whenever $s$ and $s'$ are similar. (See Fig. 6.1, where the graph of $\varphi$ is illustrated for $M = 100$). Nevertheless, "exceptional" pairs of inputs $s, s'$ for which $\sigma_{\mathcal{S}}(s,s') = 1$ and $\sigma_{\mathcal{R}}(\varphi(s),\varphi(s')) = 0$ still exist (e.g., $s = M-1$, $s' = M$). Thus, one has to take $m \equiv 0$ in order to guarantee the correctness of (6.9). Then, however, case-based inference via (6.7) becomes meaningless, since $\widehat{\varphi}_{m,\mathcal{M}}(s_0) = \mathcal{R} = \mathfrak{N}_0$ for all $s_0 \in \mathcal{S}$. □

This example suggests looking for generalized inference schemes which are less restrictive. In this chapter, we consider two possibilities of weakening the formalization of the CBI principle based on gradual rules. Firstly, we give up the requirement of its *global* validity, i.e., the fact that *one* modifier has to be determined such that (6.8) is satisfied for *all* (tuples of) cases. A related approach will be proposed in Section 6.5, where case-based inference will not be formalized by means of a single modifier, but by means of a set of ("locally valid") fuzzy rules. This idea is similar to the use of *local* similarity profiles in the constraint-based approach to CBI.

Secondly, (6.8) is obviously not very flexible in the sense that it does not allow for incorporating some tolerance toward exceptions into the inference process. In fact, the above example suggests looking for inference schemes which do not only distinguish between the possibility and impossibility of outcomes, but which are able to derive more expressive predictions using a *graded* notion of possibility. For this reason, we shall consider so-called *certainty rules* in Section 6.2 below. Replacing gradual rules by certainty rules is motivated in the same way as passing from constraint-based to probabilistic CBI as proposed in Chapter 4.

## 6.2 Certainty rules

A certainty rule corresponds to statements of the form "the more $X$ is $A$, the more *certain* $Y$ lies in $B$." More precisely, it can be interpreted as a collection of rules "if $X = x$, it is certain at least to the degree $A(x)$ that $Y$ lies in $B$"

$(x \in D_X)$, which amounts to saying that the possibility of values outside $B$ is bounded by $1 - A(x)$. This translates into the following constraint on the conditional possibility distribution $\pi_{Y|X}$ [124]:

$$\forall\, x \in D_X\,, y \in D_Y \;:\; \pi_{Y|X}(y \,|\, x) \leq \max\{1 - A(x), B(y)\}. \tag{6.15}$$

More generally, rules of this kind can be classified as *certainty-qualifying rules* [118]. The semantics of such rules is adequately captured by means of so-called S(trong)-implication operators. The latter is of the form $\alpha \rightsquigarrow \beta \overset{\mathrm{df}}{=} n(\alpha) \oplus \beta$, where $n(\cdot)$ is a strong negation and $\oplus$ a t-conorm. A special case of an S-implication is the Kleene-Dienes implication in (6.15). Note that the mapping $x \mapsto 1 - x$ in (6.15) is actually thought of as the order-reversing mapping of the ordinal scale $\mathcal{L}$.

The upper bound (6.15) implies that the possibility of $Y = y$ is bounded by $1 - A(x)$ if $X = x$ and $B(y) = 0$, which means that $y$ is outside of the support of $B$. Thus, the larger $A(x)$, the smaller the possibility that $y$ lies outside of $B$. Within the framework of possibility theory, *certainty* is closely related to *impossibility*[5] and, hence, (6.15) indeed means that $y$ lies in $B$ with certainty $A(x)$.

Since a certainty rule is thought of as a constraint which holds true in general but still allows for exceptions (see e.g. [376]), it is more flexible than the approach based on gradual rules and seems to be particularly suitable as a formal model of CBI. In connection with the concept of a certainty rule, the CBI hypothesis can be understood as "the larger the similarity of two inputs is, the more *certain* it is that the similarity of corresponding outcomes is large," an interpretation which emphasizes the heuristic nature of this assumption.

Given a new input $s_0$, an observed case $\langle s_1, r_1 \rangle \in \mathcal{M}$ constrains the possibility of similarity degrees $y = \sigma_{\mathcal{R}}(r_0, r_1)$ according to the certainty rule model (6.15):

$$\pi(y \,|\, x) \leq \pi_{cert}(x, y) = \max\{1 - y, x\}, \tag{6.16}$$

where $x = \sigma_{\mathcal{S}}(s_0, s_1)$ is the similarity between $s_0$ and $s_1$. Since $r_0 = r$ implies $y = \sigma_{\mathcal{R}}(r, r_0)$, we thus obtain

$$\pi_{s_0}(r) = \pi(r \,|\, s_0) \leq \max\,\left\{1 - \sigma_{\mathcal{S}}(s_0, s_1), \sigma_{\mathcal{R}}(r, r_1)\right\} \tag{6.17}$$

for the possibility that $r \in \mathcal{R}$ corresponds to the unknown outcome $r_0$. The more similar the inputs $s_0$ and $s_1$ are, the more constrained the possibility of outcomes becomes according to (6.17). If, for instance, $\sigma_{\mathcal{S}}(s_0, s_1)$ is close to 1, the possibility bound $\pi(r \,|\, s_0)$ can only be large for outcomes which are very similar to $r_1$. If, however, $\sigma_{\mathcal{S}}(s_0, s_1)$ is very small, we also obtain a large possibility bound for outputs hardly similar to $r_1$. Particularly, (6.17) becomes trivial if $\sigma_{\mathcal{S}}(s_0, s_1) = 0$. The resulting possibility distribution $\pi \equiv 1$ reveals *complete ignorance*. That is,

[5] Formally, the certainty $c$ of an event $A$ and the possibility $p$ of the complement of $A$ are related according to $c = 1 - p$ (cf. Section 5.1).

the observed outcome $r_1$ says nothing about the unknown outcome $r_0$, because the corresponding inputs are not similar at all.

Since (6.17) applies to all cases of the memory, we obtain the possibility distribution

$$\begin{aligned} \pi_{s_0} : r \quad &\mapsto \quad \pi(r \,|\, s_0) \\ &\overset{\mathrm{df}}{=} \min_{1 \leq \imath \leq n} \max \left\{ 1 - \sigma_{\mathcal{S}}(s_0, s_\imath), \sigma_{\mathcal{R}}(r, r_\imath) \right\}, \end{aligned} \tag{6.18}$$

which emerges from (6.15) under the application of the *minimal specificity principle*.[6] The constraint (6.18) can be generalized to

$$\begin{aligned} \pi_{s_0} : r \mapsto \pi(r \,|\, s_0) & \\ = \min_{1 \leq \imath \leq n} m_2 \left( \max \left\{ 1 - m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath) \right\} \right) & \end{aligned} \tag{6.19}$$

by means of modifier functions $m_1, m_2 : \mathcal{L} \longrightarrow \mathcal{L}$. The associated certainty rule, denoted $m_1 \circ \sigma_{\mathcal{S}} \overset{m_2}{\rightsquigarrow} \sigma_{\mathcal{R}}$, corresponds to statements of the form “for $m_1$-similar inputs it is $m_2$-certain that the respective outputs are similar.” As in the case of possibility rules, the modifier $m_2$ can be used for bounding the effect of a rule (cf. Section 5.4.4). Discounting a certainty rule can be realized, e.g., by means of a modifier $x \mapsto \max\{x, \lambda\}$, where the discounting factor $\lambda$ guarantees a minimal degree of possibility.[7]

Remark 6.2. The modifier $x \mapsto \max\{x, \lambda\}$ corresponds to a special case of the discounting operation $x \mapsto (1 - \lambda) \otimes x + \lambda$ [402]. It is obtained by taking the generalized conjunction $\otimes$ as $(\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$. The modifier $x \mapsto \min\{x, 1 - \lambda\}$, used as a discounting operation in the possibilistic framework of Chapter 5, emerges under the same conjunction from $x \mapsto (1 - \lambda) - (1 - \lambda) \otimes (1 - x)$. □

According to the gradual rule model, an observed case $\langle s_1, r_1 \rangle$ rules out the existence of other (hypothetical) cases completely, namely those which do not obey (6.8). Particularly, the set

$$\left\{ r \in \mathcal{R} \,|\, m(\sigma_{\mathcal{S}}(s_0, s_1)) \leq \sigma_{\mathcal{R}}(r, r_1) \right\}$$

of outcomes regarded as possible for the input $s_0$ excludes outputs which are not similar enough, namely those outcomes $r \in \mathcal{R}$ with $\sigma_{\mathcal{R}}(r, r_1) < m(\sigma_{\mathcal{S}}(s_0, s_1))$. As opposed to this, a certainty rule (6.17) only gradually restricts the possibility of a case $\langle s, r \rangle$:

[6] According to this principle, each element of the domain of a possibility distribution is assigned the largest possibility in agreement with the given constraints. The principle is already discussed under the name *principle of maximal possibility* in [415] and has been introduced as an information-theoretic principle in [113].

[7] This contrasts with the discounting of possibility rules, where the application of the min-operator instead of the max-operator yields an upper rather than a lower possibility bound.

$$\pi(s,r) \leq \pi_{\mathcal{C}}(s,r) = \max\left\{1 - \sigma_{\mathcal{S}}(s,s_1), \sigma_{\mathcal{R}}(r,r_1)\right\}. \tag{6.20}$$

Thus, it does generally not exclude other cases completely. In fact, the possibility of a case $\langle s, r\rangle$ is 0 only if both, $s$ is perfectly similar to $s_1$ and $r$ is completely different from $r_1$. Given a new input $s_0$, we hence obtain $\pi_{s_0}(r) > 0$ as soon as $\sigma_{\mathcal{S}}(s_0,s_1) < 1$ or $\sigma_{\mathcal{R}}(r,r_1) > 0$. It is exactly this property which allows for the modeling of exceptional inputs and which seems advantageous in connection with the adaptation of CBI models.

Example 6.3. To illustrate this, let us reconsider Example 6.1. The fact that we have to take $m \equiv 0$ in connection with the gradual rule model means that a case $\langle s, r\rangle$ no longer constrains the possibility of outcomes associated with a new input $s_0$. Now, suppose that we define $m_1$ by $m_1(0) = 0$ and $m_1(1) = 1 - \varepsilon$ (and that we take $m_2 = \mathrm{id}$) in the certainty rule approach (6.19), where $0 < \varepsilon \ll 1$. Given a case $\langle s_1, r_1\rangle$ and a new input $s_0$ similar to $s_1$, we obtain

$$\pi_{s_0}(r) = \begin{cases} 1 & \text{if} \quad \sigma_{\mathcal{R}}(r,r_1) = 1 \\ \varepsilon & \text{if} \quad \sigma_{\mathcal{R}}(r,r_1) = 0 \end{cases}. \tag{6.21}$$

Thus, outcomes which are similar to $r_1$ are regarded as completely possible, but a positive (even if small) degree of possibility is also assigned to outcomes $r$ which are not similar to $r_1$. This takes the existence of exceptional pairs of inputs into account. □

As pointed out in [99], a certainty rule (6.17) fails to modulate the width of the neighborhood around an observed outcome $r_1$ in terms of the similarity between $s_0$ and $s_1$, which a gradual rule would do. As expressed by (6.17), it only attaches a level of uncertainty (which depends on $\sigma_{\mathcal{S}}(s_0,s_1)$) to the fuzzy set $r \mapsto \sigma_{\mathcal{R}}(r,r_1)$ of outcomes close to $r_1$. A way of remedying this problem would be to use implication operators such as

$$\alpha \rightsquigarrow \beta = \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ 1-\alpha & \text{if} \quad \alpha > \beta \end{cases} \tag{6.22}$$

or

$$\alpha \rightsquigarrow \beta = \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ \max\{1-\alpha, \beta\} & \text{if} \quad \alpha > \beta \end{cases} \tag{6.23}$$

in place of $\max\{1-\alpha,\beta\}$ in (6.15).[8] Implications of that kind can be obtained from an R-implication $\rightarrow$ by contraposition, i.e., $\alpha \rightsquigarrow \beta = (1-\beta) \rightarrow (1-\alpha)$.

We then obtain the (generalized) model

$$\pi_{s_0} : r \mapsto \pi(r \,|\, s_0) = \min_{1 \leq \imath \leq n} m_2\left(m_1(\sigma_{\mathcal{S}}(s_0,s_\imath)) \rightsquigarrow \sigma_{\mathcal{R}}(r,r_\imath)\right). \tag{6.24}$$

[8] (6.23) is the $R$-implication and, at the same time, the $S$-implication related to a t-norm called the nilpotent minimum. Given a strong negation $n$, the latter is defined as $x \otimes y = \min\{x,y\}$ if $y > n(x)$ and $x \otimes y = 0$ otherwise [150].

This approach avoids the following effect which occurs under the application of the constraint (6.17): If the inputs $s_0$ and $s_1$ are similar enough, the bound of $\pi(r \,|\, s_0)$ in (6.17) only reflects the similarity between $r$ and $r_1$. This, however, means that we generally have $\pi(r \,|\, s_0) < 1$ even for outcomes $r$ which are rather similar to $r_1$. In fact, (6.17) reduces the possibility of $r_0 = r$ even if $\sigma_{\mathcal{S}}(s_0, s_1) \leq \sigma_{\mathcal{R}}(r, r_1)$. In this situation it appears to be more restrictive than a gradual rule. Observe that (6.22) to some degree combines the effect of gradual and certainty rules since $r_0 \in \sigma_{\mathcal{R}}(r_\imath, \cdot)_\alpha$ with certainty $\alpha = m_1(\sigma_{\mathcal{S}}(s_0, s_\imath))$ for all $1 \leq \imath \leq n$ (if $m_2 = \mathrm{id}$). Now, however, the certainty level and the level of the cut of the similarity relation $\sigma_{\mathcal{R}}(r_\imath, \cdot)$ are directly related (through $m_1$).

## 6.3 Cases as information sources

As in Section 4.5, we shall now look at cases as individual information sources and consider case-based inference as the parallel combination of such information sources. A corresponding (probabilistic) framework allows for a semantic interpretation of the prediction $\pi_{s_0} = \pi(\cdot \,|\, s_0)$ derived from a (modified) certainty rule. This interpretation gives a concrete meaning to a degree of possibility $\pi(r \,|\, s_0)$ and might hence be helpful in connection with the acquisition of modifiers (which act on possibility distributions). At the same time, it establishes a connection between the approaches presented in Section 6.1 and Section 6.2, showing that the latter can be seen as a generalization of the former (from a probabilistic point of view). Again, let us mention that we give up the ordinal interpretation of the underlying possibility scale in this section.

### 6.3.1 A probabilistic model

When making use of the CBI hypothesis formalized by means of a fuzzy rule, each observed case provides some evidence concerning the unknown outcome $r_0$. Given a memory $\mathcal{M}$ of $n$ cases, the individual pieces of evidence have to be combined into a global constraint. Seen from this perspective, each case serves as an information source, and one task arising in connection with CBI is the parallel combination of these information sources. In Section 6.1, for instance, the evidence derived from an individual case $\langle s_1, r_1 \rangle$ is given in the form of a set $\mathcal{N}_{m(\sigma_{\mathcal{S}}(s,s_0))}(r)$ of possible candidates, where

$$\mathcal{N}_\alpha(r_1) \overset{\mathrm{df}}{=} \left\{ r \in \mathcal{R} \,|\, \alpha \leq \sigma_{\mathcal{R}}(r, r_1) \right\}$$

denotes the *α-neighborhood* of the outcome $r_1$. Moreover, the (conjunctive) combination of evidence is realized by means of the intersection (6.9).

Recall the framework of the parallel combination of information sources which has been outlined in Section 4.5: Let $\Omega$ denote a set of alternatives, consisting of all possible states of an object under consideration and let $\omega_0 \in \Omega$ be the actual (but unknown) state. An imperfect specification of $\omega_0$ is a tuple $\Gamma = (\gamma, p_C)$, where $C$ is a (finite) set of *specification contexts*, $\gamma$ is a mapping $\gamma : C \longrightarrow 2^{\Omega}$, and $p_C$ is a probability measure over $C$. The problem of combining evidence is defined as generating an imperfect specification $\Gamma$ of $\omega_0$ which performs a synthesis among the $n$ imperfect specifications $\Gamma_1, \ldots, \Gamma_n$ issued by different information sources.

In Section 6.1, the evidence derived from an individual case $\langle s_1, r_1 \rangle$, namely the set $\mathcal{N}_{m(x)}(r_1)$ with $m(x) = m(\sigma_{\mathcal{S}}(s_0, s_1))$ being the lower similarity bound (6.11), corresponds to a particular imperfect specification $\Gamma = (\gamma, p_{C_x})$:

$$\begin{array}{lcl} C_x & = & D_{\mathcal{R}}, \\ \gamma(c) & = & \mathcal{N}_c(r_1), \\ p_{C_x}(c) & = & \left\{ \begin{array}{lll} 1 & \text{if} & c = m(x) \\ 0 & \text{if} & c \neq m(x) \end{array} \right. . \end{array} \tag{6.25}$$

A context $c$ is hence thought of as the lower similarity bound $m(x) \in D_{\mathcal{R}}$ associated with the similarity degree $x \in D_{\mathcal{S}}$. Observe that the information source $\langle s_1, r_1 \rangle$ is *correct* in the sense that the prediction $\gamma(c) = \mathcal{N}_c(r_1)$ contains the object $\omega_0 = r_0$ under the assumption that the context $c$ is true (and the modifier $m$ is admissible). It is also of *maximum specificity* since $\mathcal{N}_c(r_1)$ is the most specific characterization of $r_0$ that can be inferred by $\langle s_1, r_1 \rangle$ in this context.

The one-point distribution $p_{C_x}$ in (6.25) suggests the lower similarity bound to be known precisely. In general, however, knowledge about $m(x)$ will be incomplete. Let us therefore assume $p_{C_x}$ to be defined in a more general way, such that $p_{C_x}(c)$, the probability that $m(x) = c$, can take values between 0 and 1. Since $m(x) = c$ means that $c$ defines the (largest) lower similarity bound, it implies $\sigma_{\mathcal{R}}(r_0, r_1) \in [c, 1]$. That is, the true similarity between $r_1$ and the unknown outcome $r_0$ is at least $c$. For $y \in D_{\mathcal{R}}$, the probability that $\sigma_{\mathcal{R}}(r_0, r_1) = y$ is hence bounded as follows:

$$\mathbb{P}(y) \leq \sum_{c \in D_{\mathcal{R}} : c \leq y} p_{C_x}(c).$$

When interpreting a possibility distribution $\pi$ on $D_{\mathcal{R}}$ as an encoding of upper degrees of probability[9] – by virtue of the correspondence $\pi(y \,|\, x) = \mathbb{P}(y)$ – it is possible to trace the possibility distribution

$$\pi_{cert} : y \mapsto \pi_{cert}(y \,|\, x) = m(x) \rightsquigarrow y \tag{6.26}$$

derived from a (modified) certainty rule[10] back to a probabilistic specification of the similarity bound $m(x)$. Consider as an example (6.26) for the implication operator (6.22):

$$\pi_{cert}(y \,|\, x) = \left\{ \begin{array}{cll} 1 & \text{if} & m(x) \leq y \\ 1 - m(x) & \text{if} & m(x) > y \end{array} \right. . \tag{6.27}$$

[9] Here, we clearly give up the ordinal interpretation of the possibility scale.

[10] For the sake of simplicity, we restrict ourselves to certainty rules with one modifier in this section.

For $m(x) > 0$, (6.27) corresponds to the probability $p_{C_x}$ defined by

$$p_{C_x}(c) = \begin{cases} 1 - m(x) & \text{if} \quad c = 0 \\ m(x) & \text{if} \quad c = m(x) \\ 0 & \text{if} \quad c \notin \{0, m(x)\} \end{cases} . \tag{6.28}$$

This model can be interpreted as follows: The lower similarity bound is estimated by $m(x)$, but this estimation is only correct with a certain probability. Particularly, (6.28) assigns a positive probability to the value 0, i.e., it does not exclude the existence of outcomes which are not similar at all (and hence entail $m(x) = 0$). Associating $m(x)$ with the interval $[m(x), 1]$, we might also interpret this model as a kind of confidence interval for a similarity degree $y = \sigma_{\mathcal{R}}(r_0, r_1)$, supplemented with a corresponding level of confidence.

Since $m(x) = c$ also implies

$$r_0 \in \left\{ r \in \mathcal{R} \,|\, \sigma_{\mathcal{R}}(r, r_1) \geq c \right\},$$

the possibility distribution

$$\pi_{s_0}(r) = m(\sigma_{\mathcal{S}}(s_0, s_1)) \rightsquigarrow \sigma_{\mathcal{R}}(r, r_1), \tag{6.29}$$

which is induced by an observed case $\langle s_1, r_1 \rangle$ in connection with a certainty rule, can be interpreted in the same way as the corresponding distribution (6.26). That is, the value $\pi_{s_0}(r)$ can be interpreted as an upper bound to the probability that $r_0 = r$.

The probability (6.28) reveals a special property of the uncertain prediction derived from the rule (6.27). Namely, the certainty level associated with the estimation of a similarity bound is in direct correspondence with the similarity degree itself. That is, the larger the estimation of the similarity bound $m(x)$ is, the larger will be the level of confidence attached to the confidence interval $[m(x), 1]$.[11]

### 6.3.2 Combination of information sources

So far, we have considered only one piece of evidence, derived from a single case $\langle s_1, r_1 \rangle$, and the imperfect specification related to the corresponding similarity bound $m(x)$, where $x = \sigma_{\mathcal{S}}(s_0, s_1)$. In general, the memory $\mathcal{M}$ contains several cases, and uncertainty concerning the complete modifier (6.11) has to be specified. Thus, let us define the set of specification contexts as $C = D_{\mathcal{R}}^{D_{\mathcal{S}}}$. Each context $c \in C$ corresponds to a function $c : D_{\mathcal{S}} \longrightarrow D_{\mathcal{R}}$ and, hence, specifies a lower similarity bound $c(x)$ for all $x \in D_{\mathcal{S}}$. Moreover, suppose a certainty rule with modifier $m$ to be given and let $p_C$ be defined on $C$ in such a way that the marginal distributions correspond to the distributions $p_{C_x}$ $(x \in D_{\mathcal{S}})$ induced by this rule.

[11] Needless to say, this property is not always appropriate.

The different information sources associated with cases in the memory now share a common set $C$ of specification contexts. Let $\Gamma_\imath = (\gamma_\imath, p_C)$ $(1 \leq \imath \leq n)$ denote the imperfect specification associated with the $\imath$-th case $\langle s_\imath, r_\imath \rangle$. The mapping $\gamma_\imath$ is then given by

$$\gamma_\imath(c) = \mathcal{N}_{c(\sigma_{\mathcal{S}}(s_\imath, s_0))}(r_\imath)$$

for all $c \in C$. Making use of all cases and assuming the specification context $c \in C$ to be true, we can derive the prediction $r_0 \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)$, where

$$\widehat{\varphi}_{c,\mathcal{M}}(s_0) = \bigcap_{1 \leq \imath \leq n} \left\{ r \in \mathcal{R} \,|\, c(\sigma_{\mathcal{S}}(s_0, s_\imath)) \leq \sigma_{\mathcal{R}}(r, r_\imath) \right\}. \tag{6.30}$$

This is in accordance with the gradual rule model that considers only one modifier and, hence, provides the corresponding set-valued prediction (6.30). In fact, (6.30) reveals that each context $c \in C$ corresponds to some modified gradual rule. In other words, a certainty rule can be interpreted as a "random" gradual rule, i.e., a class of (modified) gradual rules with associated probabilities. This relation between gradual and certainty rules is further explored in Appendix B.

When considering the modifier $m$ as a random variable, the prediction of $r_0$ according to (6.30) becomes a random set, where $\widehat{\varphi}_{c,\mathcal{M}}(s_0)$ occurs with probability $p_C(c)$.[12] The probability that a certain output $r \in \mathcal{R}$ is an element of this set is given by

$$\mathbb{P}(r \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)) = \sum_{c \,:\, r \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)} p_C(c) \tag{6.31}$$

and defines an upper bound to the probability that $r_0 = r$. In connection with the idea of a randomized gradual rule model, (6.31) corresponds to the probability of selecting a (modified) gradual rule $c$ which does not exclude the (hypothetical) case $\langle s_0, r \rangle$, i.e., for which (6.30) holds.

The imperfect specification $\Gamma = (\gamma, p_C)$ defined by

$$\gamma(c) = \widehat{\varphi}_{c,\mathcal{M}}(s_0)$$

for all $c \in C$ (and $C, p_C$ as above) corresponds to the *conjunctive pooling* of the information sources $\Gamma_1, \ldots, \Gamma_n$. This kind of combination is justified by the fact that all information sources are correct with respect to all specification contexts $c \in C$. Within a possibilistic setting, conjunctive pooling comes down to deriving the intersection of possibility distributions. In fact, it is not difficult to show that (6.31) is bounded from above by the possibility distribution $\pi_{s_0}$ derived from a certainty rule in connection with a number of cases. That is,

$$\mathbb{P}(r \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)) \leq \pi_{s_0}(r) = \min \left\{ \pi^1_{s_0}(r), \ldots, \pi^n_{s_0}(r) \right\} \tag{6.32}$$

for all $r \in \mathcal{R}$, where $\pi^\imath_{s_0}$ denotes the possibility distribution derived from the $\imath$-th case according to (6.29). The interpretation of possibility degrees as upper

[12] Observe, however, that $c \neq c' \not\Rightarrow \widehat{\varphi}_{c,\mathcal{M}}(s_0) \neq \widehat{\varphi}_{c',\mathcal{M}}(s_0)$.

approximations of probabilities is hence in agreement with the application of the minimum operator in (6.19), i.e., with making use of this operator in order to combine the possibility distributions derived from individual cases.

Appendix B shows that the above probability distribution $p_C$, where $p_C(c)$ is the probability of the gradual rule associated with the context (= modifier) $c$, is unique under the assumption that the operator modeling the implication-based fuzzy rule satisfies a certain (non-)monotonicity condition. This might be considered as an interesting result, especially with regard to the combination of evidence in the probabilistic framework of Section 4.5.3. As pointed out there, the joint probability measure $\mu$ in (4.27) is generally not defined in a unique way.

According to the interpretation proposed in this section, the certainty rule approach can be seen as a generalization of the approach based on gradual rules, in the sense that the lower similarity bounds, which guarantee the correctness of the set-valued prediction of $r_0$, are no longer assumed to be precisely known. The incomplete knowledge concerning these bounds is characterized by means of a probability distribution. This allows for interpreting the case-based inference scheme in Section 6.2 as a kind of approximate probabilistic reasoning. More precisely, a prediction $\pi(\cdot \,|\, s_0)$ specifies possibility degrees $\pi(r \,|\, s_0)$ which can be seen as upper bounds to the probability that the unknown output $r_0$ is given by the outcome $r$.

## 6.4 Exceptionality and assessment of cases

Considering cases as individual information sources, as we have done in Section 6.3, suggests to rate their contribution to the prediction of outcomes. In fact, the assessment of information sources is supported by most frameworks for the combination of evidence. The basic idea, then, is to realize some kind of weighted aggregation procedure or to modify (discount) the information provided by a source according to its reliability.[13] In Section 4.6, this idea has already been discussed in the context of the probabilistic approach to CBI.

Recall that, given the same information in the form of a context $c \in C$, i.e., a modifier specifying lower similarity bounds, different cases provide different specifications of the unknown outcome $r_0$: Considering this modifier and the new input $s_0$, a case $\langle s, r \rangle$ provides a prediction of $r_0$ in the form of a possibility distribution which supports outcomes in the neighborhood of $r$. Such a specification might hence be misleading, e.g., if the outcome $r$ is rather "untypical."

Example 6.4. Consider again Example 6.1 and suppose that $s_0 = M - 1$ and $s_1 = M + 1$. In accordance with the certainty rule model (6.21) of this example

[13] See e.g. [272] for various approaches to the discounting of expert opinions within a generalized probabilistic framework.

(cf. Section 6.2), the case $\langle s_1, r_1 \rangle = \langle M+1, 1 \rangle$ strongly supports the outcomes $\{0, \ldots, 11\}$ which are similar to $r_1 = 1$. It almost rules out all other outputs, including the true outcome $r_0 = M - 1$. Loosely speaking, the (otherwise useful) information about similarity relations, specified by the certainty rule, is "misinterpreted" by $\langle s_1, r_1 \rangle$. Even though the advice to disqualify outcomes which are not similar to $r$ will lead to good predictions for the majority of cases $\langle s, r \rangle$, it is hardly reasonable when taken up in connection with an "exceptional" pair of cases, such as $\langle s_0, r_0 \rangle$ and $\langle s_1, r_1 \rangle$. □

The above example makes clear that exceptionality is not necessarily a property of an individual input or case. Rather, the label of exceptionality applies to *pairs* of cases. In fact, $\langle s_1, r_1 \rangle$ is exceptional only in connection with inputs $s = M - k$, where $1 \leq k \leq 9$, but it will lead to correct predictions for all other inputs. Moreover, the decision whether to call two cases exceptional will often not be as obvious as in our example, where only two degrees of similarity are distinguished. Making use of richer scales including intermediate degrees of similarity, exceptionality will become a gradual property.

Interestingly enough, the certainty rule framework suggests computing a degree of exceptionality in the following way:

$$\mathsf{ex}(\langle s, r \rangle, \langle s', r' \rangle) \overset{\mathrm{df}}{=} 1 - \pi_{cert}(\sigma_{\mathcal{R}}(r, r') \,|\, \sigma_{\mathcal{S}}(s, s')). \tag{6.33}$$

That is, the exceptionality of the tuple of cases $\langle s, r \rangle$, $\langle s', r' \rangle$ is inversely related to the possibility of observing $\sigma_{\mathcal{R}}(r, r')$-similar outcomes for $\sigma_{\mathcal{S}}(s, s')$-similar inputs, as specified by the certainty rule model.[14] The more $\langle s, r \rangle$ and $\langle s', r' \rangle$ violate the certainty rule, the more exceptional they are in the sense of (6.33).

It is worth mentioning that (6.33) also makes sense in connection with the gradual rule model. Applying (6.33) to the possibility distribution (6.2) induced by a gradual rule, a tuple of cases is either completely exceptional or not exceptional at all. In fact, (6.33) may also be seen as a reasonable generalization of this rather obvious definition of exceptionality. This again reveals the difference between the gradual and the certainty rule model: The former is indeed not *tolerant* toward exceptions in the sense that each violation of the rule is "punished" by classifying the involved cases as *completely* exceptional ones. As opposed to this, exceptionality is a gradual property in the certainty rule model.

Even though a gradual or certainty rule can only be violated by *tuples* of cases and, hence, exceptionality should be considered as a property of pairs of cases, it seems intuitively clear in our example that the most unreliable information sources are those cases $\langle s, r \rangle$ with $s$ close to integers $kM$ ($k \in \mathfrak{N}_0$). The closer an input is to such a point, the more likely the case might be called exceptional. In fact, one possibility of regarding exceptionality as a property of an individual

[14] Again, note that $x \mapsto 1 - x$ in (6.33) actually represents the order-reversing mapping of a possibility scale.

case $\langle s, r\rangle$ is to consider the likelihood or possibility of $\langle s, r\rangle$ to be exceptional with respect to a new case $\langle s_0, r_0\rangle$. Thus, one might think of generalizing (6.33) as follows:

$$\mathsf{ex}_1(\langle s, r\rangle) \overset{\text{df}}{=} \sup_{\langle s', r'\rangle \in \varphi} \mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle). \tag{6.34}$$

Assigning a degree of exceptionality to a case in the sense of (6.34) can be interpreted as rating the reliability of this case. Of course, this degree of exceptionality depends on the formalization of the underlying rule. In other words, a case is exceptional not by itself but only with respect to a particular rule: Changing the rule by means of a modifier also changes the degree of exceptionality of the case. For instance, the modification of a gradual rule, as proposed in Section 6.1.2, can be interpreted as adapting the rule in such way that no exceptional cases exist at all. Likewise, no case is exceptional with respect to the certainty rule in its weakest form, as formalized by $m_1 \equiv 0$ in (6.19). In connection with the certainty rule model (6.21) of Example 6.1, we obtain

$$\mathsf{ex}_1(\langle s, r\rangle) = \begin{cases} 1-\varepsilon & \text{if} \quad \exists\, k \in \mathfrak{N}_0 \,:\, |s - kM| \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

for all $\langle s, r\rangle \in \varphi$.

Let us briefly hint at two properties of (6.34). Firstly, this definition of exceptionality is completely independent of any kind of *frequency*, i.e., the value $\mathsf{ex}_1(\langle s, r\rangle)$ should not be understood as a probability of $\langle s, r\rangle$ being exceptional with respect to some other case. Of course, defining exceptionality of an individual case by using an averaging operator in place of the supremum in (6.34) seems intuitively appealing and would clearly make sense within a probabilistic setting. Recall, for instance, the probabilistic interpretation of the certainty rule model proposed in Section 6.3. According to this interpretation, a certainty rule can be seen as a collection of (modified) gradual rules to each of which is attached a certain probability. Since a case is either exceptional or not with respect to a fixed gradual rule, it is an obvious idea to derive a corresponding probability of being exceptional with respect to a certainty rule.

Secondly, (6.34) is rather strict in the sense that it implies

$$\mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle) \leq \min\{\mathsf{ex}_1(\langle s, r\rangle), \mathsf{ex}_1(\langle s', r'\rangle)\} \tag{6.35}$$

for all cases $\langle s, r\rangle$ and $\langle s', r'\rangle$. In other words, having encountered an exceptional tuple of cases, *both* cases are considered to be exceptional. This principle can obviously be weakened by concluding on the exceptionality of *at least one* of the two cases. This leads to the constraints

$$\mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle) \leq \max\{\mathsf{ex}_1(\langle s, r\rangle), \mathsf{ex}_1(\langle s', r'\rangle)\} \tag{6.36}$$

for all $\langle s, r\rangle, \langle s', r'\rangle \in \mathcal{S}$. Indeed, (6.36) will often appear more reasonable than (6.35). For instance, modifying the mapping $\varphi$ in Example (6.1) according to

$$\varphi(s) = \begin{cases} M & \text{if} \quad a \bmod M \neq 0 \\ 0 & \text{if} \quad a \bmod M = 0 \end{cases}$$

suggests to call the cases $\langle 0, 0\rangle, \langle M, 0\rangle, \langle 2M, 0\rangle \ldots$ exceptional and to consider all other cases to be (completely) normal. As opposed to this, (6.35) does not only qualify a case $\langle kM, 0\rangle$ itself as exceptional, but also all neighbored cases $\langle kM + a, M\rangle$ such that $1 \leq |a| \leq 10$.

A natural idea is to discount the information provided by a case based on its level of exceptionality. As already mentioned before, discounting a fuzzy restriction $F$ over a domain $D$ within the qualitative min-max framework amounts to modifying $F$ into $\max\{\lambda, F\}$, where $\lambda$ is a discounting factor [120]. Indeed, $F$ remains unchanged if $\lambda = 0$. As opposed to this, the modified restriction becomes trivial (and corresponds to the complete referential $D$) if the discounting is maximal ($\lambda = 1$). This approach can be applied to the result of case-based inference by identifying discounting factors with degrees of exceptionality. It amounts to computing

$$\pi(r \mid s) = \min_{1 \leq \imath \leq n} \max \left\{ \mathsf{ex}_1(\langle s_\imath, r_\imath\rangle), m(\sigma_{\mathcal{S}}(s, s_\imath)) \rightsquigarrow \sigma_{\mathcal{R}}(r, r_\imath) \right\}. \tag{6.37}$$

If exceptionality is equivalent to complete exceptionality, as in the gradual rule model, (6.37) comes down to removing the exceptional cases from the memory. Apart from that, the usual inference process is realized. In other words, (6.37) then corresponds to the gradual rule approach ($\rightsquigarrow$ is the Rescher-Gaines implication) restricted to the normal cases. When using the certainty rule model in (6.37), i.e., when modeling $\rightsquigarrow$ by implication operators such as (6.22) or (6.23), the level of uncertainty of an individual prediction is increased in accordance with the degree of exceptionality of the corresponding case. The CBI hypothesis underlying the generalized approach might then be characterized as follows: "The larger the similarity between $s$ and $s_0$ and the less exceptional the input $s$, the more certain our conclusion on the similarity between the associated outputs $r$ and $r_0$."

Interestingly enough, the modifications outlined above suggest a further way of adaptation: Not the strength of the rule is adapted to the class $\varphi$ of cases, but the influence of each case is modulated in accordance with its exceptionality relative to the (predefined) rule. In this connection, it also seems worth mentioning that assigning degrees of exceptionality to cases in such way that (6.36) is satisfied leads to an interesting problem from both, a mathematical as well as a semantical point of view. In addition to observed cases, one might think of using an (a priori) expert assessment of the exceptionality of cases (which then correspond to triples $\langle s, r, e\rangle$) in order to solve this problem, all the more since the minimization of some objective function subject to the constraints (6.36) might not guarantee a unique solution.

## 6.5 Local rules

The rule-based approaches to CBI outlined in previous sections are *local* in the sense that the information provided by different cases is processed and combined independently. They are, however, *global* in the sense that a (modified) fuzzy rule constitutes a constraint which is assumed to be globally valid. This becomes especially apparent in connection with the gradual rule approach, where an (admissible) modifier $m$ specifies (conditional) lower bounds to the similarity of outcomes which hold true for all (pairs of) cases. It has already been pointed out in Section 6.1 that this requirement often entails rather imprecise predictions, caused by the fact that admissible modifiers might not be very restrictive.

Instead of looking for a global rule, which is valid up to some exceptions – as discussed in connection with the certainty rule model in previous sections – one might weaken the principle of a gradual rule by specifying rules which are somehow "locally" valid. In this section, we follow the idea of adapting a fuzzy rule to each case of the memory more directly rather than the one of associating instantiations of a global rule with all observed cases (and perhaps discounting these instantiations in the sense of Section 6.4). This approach is quite similar to the specification of *local* similarity profiles and hypotheses in connection with the constraint-based and probabilistic approaches to CBI discussed in previous chapters. It differs, however, from the solution proposed in connection with the possibility rule model (cf. Section 5.4.6), where local rules have not been defined for individual cases, but for different (fuzzy) regions of the space of inputs.

Let us again consider the gradual rule model. The problem that global validity might lead to (local) predictions which are unnecessarily imprecise is already certified by Example 6.1. In fact, the necessity of taking $m \equiv 0$ leads to the useless predictions $\widehat{\varphi}_{m,\mathcal{M}}(s_0) = \mathfrak{N}_0$. Loosely speaking, a CBI strategy is not applicable because the hypothesis of similar inputs having similar outcomes is not globally satisfied. Still, it seems desirable to make use of the observation that the mapping $\varphi$ in the example is piecewise linear, i.e., that the CBI hypothesis is satisfied at least *locally*. One possibility of doing this is to partition the set $\mathcal{S}$ of inputs and to derive corresponding local models (cf. Section 5.4.6). In our example, the idea to partition $\mathcal{S}$ into sets of the form

$$\{kM, kM+1, \ldots, kM+(M-1)\} \qquad (k \in \mathfrak{N}_0)$$

suggests itself. However, since $\varphi$ is generally unknown, the definition of a partition will not always be obvious, all the more if $\mathcal{S}$ is non-numerical.

Here, we consider a second possibility, namely that of associating an individual (local) rule with each case of the memory. Thus, the idea is to define rules of the form "the more similar an input is to $s$, the more similar the associated outcome is to $r$" for each case $\langle s, r \rangle$ in the memory. The validity of such a (gradual) rule is already guaranteed by the (non-decreasing) modifier

$$m_{\langle s,r\rangle}(x) = \sup \left\{ h_{\langle s,r\rangle}(x') \,|\, x' \in D_{\mathcal{S}}, x' \leq x \right\}, \tag{6.38}$$

for all $x \in D_{\mathcal{S}}$, where

$$h_{\langle s,r\rangle}(x) = \inf_{\langle s',r'\rangle \in \varphi \,:\, \sigma_{\mathcal{S}}(s,s')=x} \sigma_{\mathcal{R}}(r,r'). \tag{6.39}$$

Since the infimum in (6.39) is taken over a smaller set of cases, (6.38) is obviously more restrictive than (6.11). Based on (6.38), the inference scheme (6.9) can be replaced by

$$r_0 \in \bigcap_{1 \leq \imath \leq n} \left\{ r \in \mathcal{R} \,|\, m_{\langle s_\imath, r_\imath\rangle}(\sigma_{\mathcal{S}}(s_0, s_\imath)) \leq \sigma_{\mathcal{R}}(r, r_\imath) \right\}. \tag{6.40}$$

In our example, the maximally constraining (admissible) modifier for a case $\langle s,r\rangle = \langle s, \varphi(s)\rangle$ is simply given by

$$m_{\langle s,r\rangle}(x) = \begin{cases} x & \text{if} \quad 10 \leq s \bmod M \leq M-9 \\ 0 & \text{otherwise} \end{cases}.$$

Based on a sufficiently large number of observations, the mapping $\varphi$ can hence be approximated rather accurately. More precisely, the prediction (6.40) converges toward

$$\widehat{\varphi}(s_0) = \begin{cases} \{\varphi(s_0), \ldots, 20\} & \text{if} \quad 0 \leq \varphi(s_0) < 20 \\ \{\varphi(s_0)\} & \text{if} \quad 20 \leq s_0\varphi(s_0) < M-20 \\ \{2M - 12 - \varphi(s_0), \ldots, M+9\} & \text{if} \quad M-20 \leq \varphi(s_0) < M \end{cases}$$

with an increasing number of observations.

Observe that a local rule can be taken as an indication of the (prediction) quality of a case $\langle s,r\rangle$ and can hence support the design of an optimal case base. The more restrictive a rule can be made by means of a modifier $m_{\langle s,r\rangle}$, the more it will contribute to precise predictions. As in our example, good local rules will generally be provided by "typical" cases, the outcomes of which are at least to some degree representative of similar inputs. In this sense, a modifier can also be seen as an assessment of a case (cf. Section 6.4). A modifier $m_{\langle s,r\rangle} < \mathrm{id}$, for instance, brings the discounting of a case about, whereas a modifier $m_{\langle s,r\rangle} > \mathrm{id}$ produces the opposite effect. Particularly, letting $m_{\langle s,r\rangle} \equiv 0$ comes down to leaving the corresponding case out of account, i.e., to remove it from the memory.

Let us mention that a (globally admissible) gradual rule can be seen as a collection of rules

$$\begin{aligned} \alpha(x) :\; & \sigma_{\mathcal{S}}(s_1, s_2) = x \;\Rightarrow \\ & \qquad \forall\, r_1 \in \varphi(s_1)\, \forall\, r_2 \in \varphi(s_2) \;:\; \sigma_{\mathcal{R}}(r_1, r_2) \geq m(x), \end{aligned}$$

each of which is an aggregation of more specific (local) rules [115] associated with cases $\langle s, r\rangle \in \varphi$. More precisely, a rule $\alpha(x)$ can be seen as an approximation in the form of a disjunction

$$\alpha(x) = \bigvee_{\langle s,r\rangle \in \varphi} \alpha(\langle s, r\rangle, x) \tag{6.41}$$

of local rules

$$\begin{aligned} \alpha(\langle s, r\rangle, x) :\ (\langle s_1, r_1\rangle = \langle s, r\rangle) \wedge (\sigma_{\mathcal{S}}(s_1, s_2) = x) \Rightarrow \\ \forall\, r_2 \in \varphi(s_2) \,:\, \sigma_{\mathcal{R}}(r, r_2) \in [m_{\langle s,r\rangle}(x), 1]. \end{aligned} \tag{6.42}$$

Since the disjunction in (6.41) is taken over all cases $\langle s, r\rangle \in \varphi$, the global rule $\alpha(x)$ depends on the similarity degree alone. Observe that (6.11) and (6.38) are related through

$$\forall\, x \in D_{\mathcal{S}} \,:\, m(x) = \inf_{\langle s,r\rangle \in \varphi} m_{\langle s,r\rangle}(x),$$

which shows that taking the disjunction of the consequent parts in (6.42) comes down to bounding similarity degrees from below and which again reveals the restrictive nature of the gradual rule model.

Interestingly enough, a certainty rule can be seen as a more general fusion of local rules (6.42), taking into account that some conclusions might be less plausible (or might occur less often) than others and, hence, may lead to a *weighted* union of conclusions instead of a disjunction.

Let us finally mention that the idea of adapting a rule-based formalization of the CBI hypothesis to individual cases applies to certainty rules in the same way as to gradual rules. Observe that local certainty rules can be seen as a combination of the two aforementioned generalizations of the gradual rule model. In fact, these rules are local and tolerant toward exceptions at the same time.

## 6.6 Summary and remarks

### Summary

- The objective of this chapter was to elaborate in more detail on implication-based fuzzy rules as an alternative model of the inference process in case-based reasoning. It has been shown that this type of rule leads to an approach which deviates considerably from the possibility rule model discussed in Chapter 5. In fact, implication-based fuzzy rules realize a *constraint-based* approach in much the same way as the method proposed in Chapter 3: Already encountered cases are looked at as evidence for (partially) ruling out other (hypothetical) cases, not similar enough to the observed ones. As opposed to this, a possibility rule is a *conjunction-based* rule and gives rise to an *example-oriented* approach:

Observed cases are considered as pieces of data which provide evidence for the possibility of observing similar cases.

- We have distinguished between two types of implication-based rules. The first type (gradual rules) assumes a kind of closeness relation between the similarity of inputs and the similarity of outcomes which is not tolerant toward exceptions. Given a new input, the observed cases which constitute the memory are taken as evidence for either allowing or completely excluding certain outcomes. A second type of rules (certainty rules) only uses case-based information for deriving conclusions about the *possibility* of outcomes. They are more expressive and allow for the *partial* exclusion of outputs. Moreover, they can formalize situations in which the CBI hypothesis holds true "in general" up to some exceptions to the "similar inputs-similar outputs" rule.
- The use of modifier functions has been proposed for modulating the "strength" of fuzzy rules. This way, it becomes possible to adapt the formal model according to the extent to which the CBI hypothesis actually holds true for the respective application.
- The meaning of exceptionality of cases has been discussed in connection with the idea of discounting cases which might be seen as somewhat unreliable or misleading information sources. The discounting of cases, in conjunction with a modification of the basic inference scheme, presents a further possibility of model adaptation.
- Local rules have been introduced as a second direction of generalizing the basic model. There are different motivations for this step: In the gradual rule model, it is true that the instantiation of a (globally) admissible rule by different cases leads to correct predictions. However, inference results might be poor since this rule will often hardly be constraining. In the certainty rule model, the multiple instantiation of the same global rule leads to difficulties in connection with exceptional (still not discounted) cases. This might cause inconsistencies and an exaggerated exclusion of (rather possible) cases. We have also pointed out a close relation between local rules and the assessment of cases. In fact, the determination of a modifier for an individual case can be seen as a rating of the typicality or prediction quality of that case. Particularly, a modifier can make a local rule completely ineffective, which amounts to removing the corresponding case from the memory. Next to the idea of exceptionality with respect to a global rule, the concept of local rules thus presents a further possibility of rating and discounting cases.

## Remarks

- In this chapter, we have refrained from discussing several issues which have already been considered in connection with the possibility rule model in Chapter 5. This concerns especially the extensions of the basic model, discussed in

Sections 5.3 and 5.4. These techniques can as well be applied to the CBI model which proceeds from implication-based fuzzy rules.

- The combination of possibility and certainty rules has already been proposed as a basis of the calibration method in Section 5.6. Besides, there are other motivations for using implication-based and conjunction-based rules jointly. In [205], for instance, it is argued that a combination of the two types of rules can greatly improve the informational contents of (possibilistic) case-based predictions. In fact, as already pointed out in Section 5.3.3, the degree $\delta_{s_0}(r)$ derived from a possibility rule can be seen as a degree of *confirmation* of the outcome $r$ and actually defines a *lower* possibility bound. As opposed to this, the degree $\pi_{s_0}(r)$ obtained in connection with a certainty rule model reflects the degree to which past experience (in the form of the memory $\mathcal{M}$) *excludes* the output $r$ and determines an *upper* degree of possibility. Recall the following extreme examples from Section 5.3.3:

  (a) $\delta_{s_0}(r) = 0$, $\pi_{s_0}(r) = 1$: A situation of complete ignorance. Neither is $r$ supported nor (partly) excluded by any observation. Thus, $r$ is fully plausible though not confirmed at all.

  (b) $\delta_{s_0}(r) = 0$, $\pi_{s_0}(r) = 0$: Clear evidence against $r$ has been accumulated in the form of inputs similar to $s_0$ with outputs dissimilar to $r$.

  (c) $\delta_{s_0}(r) = 1$, $\pi_{s_0}(r) = 1$: The output $r$ is strongly supported through the observation of similar cases.[15]

  The above cases emphasize the advantage of the combined approach. The example-based (possibility rule) model alone cannot distinguish between (a) and (b). It goes without saying, however, that it makes a great difference from an epistemic point of view whether a case is not supported simply because no similar cases have been observed or whether indeed some evidence against this case has been accumulated (through the certainty-rule model of the CBI principle). The constraint-based model cannot distinguish between the cases (a) and (c). Again, however, it might be important to know whether an outcome $r$ seems completely possible for the input $s_0$ only because no input has been observed which is similar to $s_0$ or whether $r$ is indeed supported by the observation of cases $\langle s, r \rangle$ such that $s$ is similar to $s_0$ (which requires $\pi_{s_0}(r) > 0$).

[15] In fact, a possibility degree of 1 requires the observation of a *perfectly* similar case. If the similarity relations are separating, this means that $\langle s, r \rangle$ itself has already been encountered.

# 7. Case-Based Decision Making

Early work in AI has mainly focused on formal logic as a basis of knowledge representation and has largely rejected approaches from (statistical) decision theory as being intractable and inadequate for expressing the rich structure of (human) knowledge [193]. However, the recent development of more tractable and expressive decision-theoretic frameworks and inference strategies such as, e.g., graphical formalisms [292, 187], in combination with the analysis of restrictions of traditional AI reasoning techniques have stimulated renewed interest in decision theory. In fact, ideas from decision theory now play a predominant role in the modeling of *rationality*, one of the major topics of current research in AI [95]. Loosely speaking, the AI paradigm has undergone a shift from "acting logically" to "acting rationally" [322]. The related view of intelligent behavior deviates fundamentally from the classical "logicist" approach. While the latter emphasizes the ability to reach correct conclusions from correct premises, the decision-theoretic approach considers AI as the design of (limited) *rational agents* [324]. For this "agent-based" view of AI, intelligence is strongly related to the capacity of successful behavior in complex and uncertain environments and, hence, to *rational decision making.*[1]

Decision theory and AI can fertilize each other in various ways [298]. As already suggested above, classical decision theory provides AI with important ideas and concepts of rationality, thus contributing to a formal basis of intelligent agent design. Yet, it has been less concerned with computational and knowledge representational aspects. AI can particularly contribute in this direction. It has been realized very soon, for instance, that *perfect rationality*, in the sense of generating behavior which leads to maximal (expected) utility, cannot be achieved once computational aspects come into play [321]. In fact, an agent having to make a decision under limited computational (time, memory) resources not only has to reason about the decision itself but also about the computations it uses for deriving the decision: A more elaborated computation might yield a better decision but also requires more time (or other resources). Being perfectly rational in the aforementioned sense, it has to perform the reasoning about its computations in the same decision-theoretic way. This, however, leads to the problem of realizing some kind of *metalevel rationality* [23, 38, 324] and, hence, results in a conceptual regress. Problems of this kind have motivated the definition of alternative

[1] J. Doyle has suggested to define AI itself as the computational study of rational behavior [94].

concepts which are weakenings of perfect rationality. They serve as candidates for putting the agent-based understanding of intelligence and the related approach to the design of intelligent systems on a formal basis. Among the proposals, the concept of *bounded optimality* seems to be the one which is most relevant for practical as well as theoretical AI research [323].

As far as the aspect of knowledge representation is concerned, research in AI has shown various possibilities of extending the decision-theoretic frameworks usually considered in classical approaches. Recent developments include the modeling of decision problems within qualitative [52, 53, 123, 129] and constraint-based [143] settings and make use of formal logic in order to represent the knowledge of a decision maker in a more flexible way [39, 50, 98, 293, 326, 365, 366]. These approaches are intended to make decision-theoretic models more realistic, tractable and expressive.

In this chapter, we are mainly concerned with the idea of *case-based decision making* (CBDM) which is originally due to Gilboa and Schmeidler [167]. The notion CBDM stands for the application of the CBI principle in the context of decision making: An agent faced with a decision problem relies upon its experience from similar problems encountered in the past. Loosely speaking, it chooses an act based on the (cumulative or average) performance of (potential) acts in previous problems which are similar to the current one.

Even though the model in [167] has mainly been introduced with economic applications in mind, CBDM is particularly interesting from an AI perspective. Firstly, it combines principles from two important subfields of AI, namely decision theory and CBR. Secondly, it touches on interesting aspects of knowledge representation and reasoning. In fact, the mental notions of *preference* and *belief* constitute the main concepts of classical decision theories. Corresponding mathematical models are based on formalizations of these concepts, such as preference relations, utility functions, and probability distributions. The aforementioned approach of Gilboa and Schmeidler leads to a decision theory in which the cognitive concept of *similarity* plays a predominant role. Needless to say, incorporating this concept into formal approaches to decision making raises some interesting (semantical) questions. Particularly, it has to be clarified which role similarity plays and, hence, what the relation between this and other concepts such as preference and belief could be (cf. Section 7.6). Clearly, this question concerns basic assumptions underlying a decision-theoretic model. One should therefore not expect to find definite answers. Classical works by Ramsey [309], de Finetti [146], von Neumann and Morgenstern [278] and Savage [331] as well as recent developments in the field of decision theory, such as non-additive expected utility [334, 166] or qualitative decision making, show various ways of formalizing the notions of preference and belief (including measure-theoretic approaches, such as fuzzy measures [384] and different types of probability [145], as well as more logic-oriented symbolic methods [365]).[2] Moreover, a consensus concerning the actual meaning of the concept itself seems to exist even less in the case of similarity than in the case of preference or uncertainty. As will be seen, the approaches to case-based decision making discussed in this chapter not do only differ with respect to the mathematical formalization, they are also based on different principles and ideas for incorporating similarity and principles of CBI into decision making.

The remaining part of the chapter is organized as follows: In Section 7.1, we provide a brief review and discussion of case-based decision theory as introduced by Gilboa and Schmeidler. In Section 7.2, we consider the idea of case-based decision making in connection with the Nearest Neighbor principle which is commonly used in instance-based learning. A fuzzy set-based approach to CBDM which is due to Dubois and Prade [101] will be discussed in Section 7.3. A generalization of the latter is proposed in Section 7.4. Section 7.5 is devoted to an alternative framework of case-based decision making which is based on the methods of case-based inference proposed in previous chapters. A discussion of some selected aspects of CBDM models follows in Section 7.6. Finally, Section 7.7 introduces a framework of *experienced-based decision making* as a generalization of case-based decision making.

## 7.1 Case-based decision theory

This section gives a brief review of the model introduced by Gilboa and Schmeidler [167], referred to as *case-based decision theory* (CBDT) by the authors. Putting it in a nutshell, the setup they proceed from can be characterized as follows: Let $\mathcal{Q}$ and $\mathcal{A}$ be (finite) sets of problems and acts, respectively, and denote by $\mathcal{R}$ a set of outcomes (outputs) or results. Choosing act $a \in \mathcal{A}$ for solving problem $p \in \mathcal{Q}$ leads to the outcome $r = r(p, a) \in \mathcal{R}$. A utility function $u : \mathcal{R} \longrightarrow U$ resp. $u : \mathcal{Q} \times \mathcal{A} \longrightarrow U$ assigns utility values to such outcomes; the utility scale $U$ is taken as the set of real numbers. Let

$$\sigma_{\mathcal{Q}} : \mathcal{Q} \times \mathcal{Q} \longrightarrow [0, 1], \quad \sigma_{\mathcal{R}} : \mathcal{R} \times \mathcal{R} \longrightarrow [0, 1]$$

be similarity measures quantifying the similarity of problems and outputs, respectively. Suppose the decision making agent to have a (finite) memory

$$\mathcal{M} = \{(p_1, a_1, r_1), \ldots, (p_n, a_n, r_n)\} \tag{7.1}$$

of cases at its disposal, where $(p_k, a_k) \in \mathcal{Q} \times \mathcal{A}$, $r_k = r(p_k, a_k)$ $(1 \leq k \leq n)$, and suppose furthermore that it has to choose an act for a new problem $p_0 \in \mathcal{Q}$. If a certain act $a_0 \in \mathcal{A}$ has not been applied to the problem $p_0$ so far (i.e.,

[2] Needless to say, a validation or comparison of decision-theoretic models is generally difficult, no matter whether from a descriptive or a normative point of view.

there is no case $(p_0, a_0, r) \in \mathcal{M}$) the agent will generally be uncertain about the result $r(p_0, a_0)$ and, hence, about the utility $u(r(p_0, a_0))$. According to the assumption underlying the paradigm of CBDT it then evaluates an act based on its performance in similar problems in the past, as represented by (parts of) the memory $\mathcal{M}$. More precisely, the decision maker is supposed to choose an act which maximizes a linear combination of the benefits experienced so far:

$$V(a_0) = V_{p_0,\mathcal{M}}(a_0) \overset{\text{df}}{=} \sum_{(p,a_0,r)\in\mathcal{M}} \sigma_{\mathcal{Q}}(p, p_0) \cdot u(r). \tag{7.2}$$

The summation over an empty set yields the "default value" 0 which plays the role of an "aspiration level." Despite the formal resemblance between (7.2) and the well-known expected utility formula one should not ignore some substantial differences between CBDT and expected utility theory (EUT). This concerns not only the conceptual level but also mathematical aspects. Particularly, it should be noted that the similarity weights in (7.2) do not necessarily sum up to 1. Consequently, (7.2) must not be interpreted as an estimation of the utility $u(r(p_0, a_0))$.

As an alternative to the linear functional (7.2), an "averaged similarity" version has been proposed. It results from replacing $\sigma_{\mathcal{Q}}$ in (7.2) by the similarity measure

$$(p, p_0) \mapsto \sigma_{\mathcal{Q}}(p, p_0) \left( \sum_{(p',a_0,r')\in\mathcal{M}} \sigma_{\mathcal{Q}}(p', p_0) \right)^{-1} \tag{7.3}$$

whenever the latter is well-defined. (Note that this measure is defined separately for each act $a_0$.) Theoretical details of CBDT including an axiomatic characterization of decision principle (7.2) are presented in [167].

The basic model has been generalized with respect to several aspects. The problem of optimizing decision behavior by adjusting the aspiration level in the context of repeated problem solving is considered in [168] (see also Section 7.6). In [169], the similarity measure in (7.2) is extended to problem–act tuples: Given two similar problems, it is assumed that similar outcomes are obtained for *similar* acts (not only for the same act). Indeed, it is argued convincingly that a model of the form

$$V(a_0) = \sum_{(p,a,r)\in\mathcal{M}} \sigma_{\mathcal{Q}\times\mathcal{A}}((p, a), (p_0, a_0)) \cdot u(r), \tag{7.4}$$

where $\sigma_{\mathcal{Q}\times\mathcal{A}}$ is a (problem–act) similarity measure over $\mathcal{Q} \times \mathcal{A}$, is more realistic than (7.2). For example, an act $a_0$ which has not been applied as yet is generally not evaluated by the default utility 0 if experiences with a comparable act $a$ have been made. In fact, an outcome $r(p, a)$ will then influence the rating of $a_0$ in connection with a problem $p_0$ which is similar to $p$. Besides, it should be noticed that (7.4) allows for realizing some kind of analogical reasoning. Suppose, for instance, that the effect expected from applying $a_0$ to $p_0$ is comparable to the effect of applying $a$ to $p$. In that sense, $(a_0, p_0)$ might appear to be quite similar to $(a, p)$, although $a$ and $a_0$ as well as $p$ and $p_0$ as such are rather dissimilar.

With regard to alternative models of CBDM proposed in subsequent sections it is useful to picture again the following properties of the decision criteria outlined above:

– Accumulation/averaging: The criteria (7.2) and (7.4) realize a simple summation of (weighted) degrees of utility. Consequently, a decision maker might prefer an act $a$, which always brought about rather poor results, to an act $a^*$ which has so far yielded very good results, simply because $a$ has been tried more often than $a^*$. This effect is annulled by (7.3), where the use of a normalized similarity measure yields an average utility.

– Compensation: Both decision rules compensate between good results and bad results associated with an act $a$.

Gilboa and Schmeidler especially emphasize the cognitive plausibility of their model [171]. In fact, a main motivation behind CBDT is to provide a more faithful description of human decision making than EUT does. Indeed, in some situations this axiomatic theory seems rather restrictive. Particularly, it assumes the decision maker to have very detailed information at its disposal: a list of the *states of nature*, a probability distribution over these states, a list of potential acts, and a numerical utility value for all act–state pairs.[3] Since this information is generally not completely available, the decision maker is forced to engage in *hypothetical reasoning*.[4] Moreover, some well-known paradoxes [13, 140] as well as psychological studies [375] show that EUT can be challenged as a *descriptive* theory of (human) decision making. Still, it deserves mentioning that CBDT is not seen as a competing theory, but rather as an alternative (or complementary) "language" for modeling decision problems. It seems especially useful if a problem description is not naturally cast in the framework of decision making under risk or if a problem is very unfamiliar, in which case the modeling of states of nature and associated probabilities might be difficult. A thorough discussion of the relation between CBDT and EUT can again be found in [167].

Let us conclude with a remark on the concept of similarity as used in CBDT. One might argue that the measures $\sigma_{\mathcal{Q}}$ and $\sigma_{\mathcal{Q}\times\mathcal{A}}$ need not be interpreted as similarities at all: Basically, the valuation (7.2) can be seen as a weighted sum

$$V(a_0) = V_{p_0,\mathcal{M}}(a_0) = \sum_{(p,a,r)\in\mathcal{M}} \omega_{p_0,a_0,\mathcal{M}}(p,a) \cdot u(r) \tag{7.5}$$

of utility degrees encountered in the past,[5] where the weights reflect the *relevance* of a case. This relevance, however, might not only depend on similarity. Rather,

[3] Still, it has to be noticed that an unequivocal model does generally not exist. Rather, there is much freedom in the definition of, e.g., states and acts.

[4] What is the effect of choosing a certain act in a certain state of nature?

[5] The linearity of the representation (7.2) is mainly due to the separability axiom in [167].

it can capture other (or further) aspects as well and, hence, leaves much freedom for different types of cognitive interpretation.[6] In this connection, it is worth mentioning that the axiomatic frameworks in [167, 169] do not impose special restrictions (such as symmetry) on $\sigma_{\mathcal{Q}}$ and $\sigma_{\mathcal{Q}\times\mathcal{A}}$ which might appear natural when interpreting the latter as similarity measures.

The indexing of a weight $\omega_{p_0,a_0,\mathcal{M}}(p,a)$ in (7.5) suggests that the relevance of a case $(p,a,r)$ is not necessarily a function of $(p,a)$ and $(p_0,a_0)$ alone but might also depend on other cases in the memory $\mathcal{M}$. An example of this type of "context-sensitive" relevance will be presented in the next section.

## 7.2 Nearest Neighbor decisions

Interestingly enough, the modification (7.3) of decision criterion (7.2) corresponds to a special version of a $k$-Nearest Neighbor approximation, namely Shephard's interpolation method which makes use of the complete set of observations [340]. It is used for making predictions in other CBI approaches as well (e.g., in the ELEM2-CBR system [61]). Indeed, case-based decision making can basically be seen as a special type of CBI or, more specifically, of case-based inference as discussed in previous chapters: Evaluating the act $a_0$ comes down to estimating the associated *outcome* $r(p_0,a_0)$ (resp. the utility thereof) when viewing a problem–act tuple $(p_0,a_0)$ as an *input* in the sense of CBI. In this sense, a single decision problem gives rise to several CBI problems since a corresponding estimation has to be derived for all acts $a \in \mathcal{A}$. Of course, the estimation of an outcome can principally be realized by any method of instance-based prediction.[7] In particular, one might think of replacing (7.2) by the NN rule in its basic form, an idea that we shall discuss below.

### 7.2.1 Nearest Neighbor classification and decision making

Recall the problem–act similarity model (7.4) and let $\sigma_{\mathcal{S}} = \sigma_{\mathcal{Q}\times\mathcal{A}}$ denote a similarity measure over the set of inputs which now corresponds to the set $\mathcal{Q}\times\mathcal{A}$ of problem–act tuples. Moreover, let $\mathcal{M}^{\downarrow}$ be the projection of the memory $\mathcal{M}$ to $\mathcal{Q}\times\mathcal{A}$. The NN-based counterpart to the evaluation (7.4) of an act $a_0 \in \mathcal{A}$ is then given by

$$V(a_0) = u(r(\mathsf{NN}_{\mathcal{M}}(p_0,a_0))), \tag{7.6}$$

where $\mathsf{NN}_{\mathcal{M}}(p_0,a_0)$ is the nearest neighbor of the problem–act tuple $(p_0,a_0)$:

$$\mathsf{NN}_{\mathcal{M}}(p_0,a_0) = \arg\max_{(p,a)\in\mathcal{M}^{\downarrow}} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)). \tag{7.7}$$

[6] Gilboa and Schmeidler fully agree in this point. See [71] for a related discussion and [36] for an application of CBDT where the notion of "relevance" might be preferred to that of "similarity."

[7] In fact, other machine learning methods could be used as well (cf. Section 7.7).

Of course, definition (7.7) should be refined in order to handle the non-uniqueness of the nearest neighbor. However, for the sake of simplicity we assume that each problem–act tuple $(p_0, a_0)$ has a unique nearest neighbor in $\mathcal{M}^{\downarrow}$ (according to the similarity $\sigma_{\mathcal{Q}\times\mathcal{A}}$).

Observe that the CBDT criteria (7.2) and (7.4) use all cases in order to evaluate an act. As opposed to this, the decision maker concentrates completely on the most relevant experience when evaluating an act according to (7.6). More precisely, (7.6) corresponds to (7.5) with the relevance given by

$$\omega_{p_0,a_0,\mathcal{M}}(p,a) = \begin{cases} 1 & \text{if } (p,a) = \mathsf{NN}_{\mathcal{M}}(p_0,a_0) \\ 0 & \text{otherwise} \end{cases} .$$

On the one hand, some information is clearly lost by reducing the number of cases taken into account.[8] On the other hand, the nearest neighbor does generally provide the most relevant information, i.e., the loss of information is limited.[9] Moreover, (7.6) can be seen as an approximation of (7.4) which appears reasonable from a computational point of view. Indeed, since the retrieving of all previous cases might be very time consuming, a decision maker will generally not fall back on its entire experience when having to perform a prompt action. Besides, (7.6) might appear more natural in some situations since it avoids the accumulation and compensation effect produced by (7.2) and (7.4) (cf. Section 7.1). Particularly, the estimation (7.6) corresponds to the true utility if $a_0$ has already been applied to $p_0$ in the past (which means that $(p_0, a_0) \in \mathcal{M}^{\downarrow}$). The addition of further (weighted) utility degrees or any kind of averaging might then be counterproductive (cf. Section 7.6).

Note that the NN-decision rule (7.6) partitions the set $\mathcal{A}$ into equivalence classes $[a]$, where

$$b \in [a] \;\Leftrightarrow\; a \sim b \;\Leftrightarrow\; \mathsf{NN}_{\mathcal{M}}(p_0,a) = \mathsf{NN}_{\mathcal{M}}(p_0,b).$$

In fact, two acts $a$ and $b$ are rated equally in the sense of (7.6) whenever $a \sim b$, i.e., as soon as both acts have the same nearest neighbor (in connection with a problem $p_0$). The criterion (7.6) hence ignores the actual degrees of similarity, a problem already mentioned in connection with the comparison of instance-based and kernel-based extrapolation of case-based information (cf. Section 5.3.5). This, however, does not appear reasonable from a decision making point of view. Consider, for instance, a case $(p, a, r)$ with high utility $u(r)$. Moreover, let $b$ and $c$ be acts such that $\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,b))$ is large and $\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,c))$ is small. Still, assume that $\mathsf{NN}_{\mathcal{M}}(p_0,b) = \mathsf{NN}_{\mathcal{M}}(p_0,c) = (p,a)$. In this situation, a risk-averse decision maker will generally prefer $b$ to $c$. The criterion (7.6), however, does not differentiate between these two acts. The NN principle (as any other estimation method) seems hence questionable in the context of decision making.

[8] Though formally only the relevance of some cases is set to 0.

[9] This claim can be proved in a formal way. The result in [74], for instance, can be interpreted as follows: Under certain technical assumptions, at least *half* of the information that a complete random sample contains about an outcome in already represented by the nearest neighbor of the query instance.

Indeed, at this point one should realize an important difference between decision making and *prediction*, the performance task which is commonly solved by NN algorithms: In a prediction problem, an estimation has to be derived for only *one* instance and this estimation is not considered as a valuation which supports any kind of comparison. Having to choose one among the potential candidates anyway, it might then be acceptable to base an estimation on the nearest neighbor even if it turns out to be quite dissimilar.

Let us mention that the averaged similarity criterion (7.3) suffers from a similar problem. In fact, it is readily seen that the valuation of an act according to (7.3) can be very large even though this act has only been applied in situations which are hardly similar to the current problem.

### 7.2.2 Nearest Neighbor decision rules

In order to overcome the aforementioned problem it seems natural to not only associate the utility $v$ of the nearest neighbor $(p,a) \in \mathcal{M}^{\downarrow}$ with each act $a_0 \in \mathcal{A}$ (i.e., with the tuple $(p_0,a_0)$), but rather the tuple $(v,\sigma)$, where $\sigma$ denotes the similarity between $(p_0,a_0)$ and $(p,a)$. The preferences of an agent should then be expressed in terms of a preference relation over the class of such utility–similarity tuples. This is somewhat comparable to generalized decision rules which take not only the expected utility into account but also the variance (i.e. uncertainty) related to an act.

More specifically, one might think of the following generalization of (7.6):

$$V(a_0) = \sigma_{\mathcal{Q}\times\mathcal{A}}\left((p_0,a_0), \mathsf{NN}_{\mathcal{M}}(p_0,a_0)\right) \cdot u(r(\mathsf{NN}_{\mathcal{M}}(p_0,a_0)))\,. \tag{7.8}$$

This valuation, which represents a preference relation over the set of tuples $(v,\sigma)$ by means of

$$(v,\sigma) \preceq (v',\sigma') \;\Leftrightarrow\; v\cdot\sigma \leq v'\cdot\sigma',$$

combines (7.4) and (7.6) to some extent. Again, it considers only one previous case (namely the nearest neighbor) rather than all cases when evaluating an act. The corresponding utility, however, is now weighted by the degree of similarity. In fact, (7.8) can be seen as a special version of (7.4) when interpreting $\sigma_{\mathcal{Q}\times\mathcal{A}}$ as a measure of relevance (cf. Section 7.1), which is then given by

$$\omega_{p_0,a_0,\mathcal{M}}(p,a) = \left\{ \begin{array}{cl} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)) & \text{if } (p,a) = \mathsf{NN}_{\mathcal{M}}(p_0,a_0) \\ 0 & \text{otherwise} \end{array} \right. . \tag{7.9}$$

According to (7.9), only the nearest neighbor is considered as a relevant observation. Of course, this idea might be generalized by taking the $k \geq 1$ nearest neighbors into account, or by introducing a threshold such that the relevance of an observation is set to 0 in case its similarity is too small.

The valuation (7.8) defines a reasonable tradeoff between the *goodness* (in terms of utility) and the *relevance* (in terms of similarity) of an experience. Still, it deserves mentioning that the degree of similarity is nothing else than a heuristic indication of the actual degree of uncertainty of an NN estimation. In fact, it is not true in general that a larger similarity comes along with a higher precision of an estimation.

It has already been mentioned that a reduction of observations as realized by (7.8) might be reasonable from a computational point of view. Particularly, this is true if the decision maker has a large memory of cases but a relatively small number of acts (and if it disposes of an efficient method of case retrieval). In the reverse case where the memory is small and the set of acts to be evaluated is large, a different strategy which passes through the set of cases, $\mathcal{M}$, rather than the set of acts, $\mathcal{A}$, might be preferred: Instead of considering the most relevant observation for each act one can proceed from an observation and attach the related experience to the most relevant act. This idea is realized by the following counterpart to (7.8):

$$V(a_0) = \sum_{(p,a)\in\mathcal{M}^{\downarrow}:\mathsf{NN}_{p_0,\mathcal{A}}(p,a)=a_0} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)) \cdot u(r(p,a)). \tag{7.10}$$

Here,

$$\mathsf{NN}_{p_0,\mathcal{A}}(p,a) = \arg\max_{a_0\in\mathcal{A}} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)) \tag{7.11}$$

denotes the problem–act tuple $(p_0, a_0) \in \{p_0\} \times \mathcal{A}$ which is maximally similar to the observation $(p,a) \in \mathcal{M}^{\downarrow}$. We assume (7.11) to be unique whenever some $a_0 \in \mathcal{A}$ exists such that $\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)) > 0$; otherwise we let $\mathsf{NN}_{p_0,\mathcal{A}}(p,a) = \emptyset$ by definition.

### 7.2.3 An axiomatic characterization

In [169], an axiomatization of (7.4) is proposed which assumes a preference relation $\succeq_x \subset \mathcal{A} \times \mathcal{A}$ over the set of acts to be given. As suggested by the attached index, this preference relation depends on the experience of the decision maker: $x$ defines a $\mathcal{M}^{\downarrow} \longrightarrow \mathfrak{R}^n$ function which assigns utility degrees to problem–act pairs. It can simply be thought of as the vector

$$x = (x_1, \ldots, x_n) = \big(u(r(p_1,a_1)), \ldots, u(r(p_n,a_n))\big)\,,$$

where $x_\imath = u(r(p_\imath, a_\imath)) \in \mathfrak{R}$ corresponds to the utility obtained in connection with the $\imath$-th problem–act tuple $(p_\imath, a_\imath)$. The vector $x$ represents the history of the decision maker and determines the *context* of the new decision problem. The information available to a decision maker which has to evaluate an act $a \in \mathcal{A}$ might thus be illustrated in the form of a table as follows:

| | | | | |
|---|---|---|---|---|
| utility | $x_1$ | $x_2$ | $\ldots$ | $x_n$ |
| similarity | $\sigma_1(a)$ | $\sigma_2(a)$ | $\ldots$ | $\sigma_n(a)$ |

(7.12)

The (case-based) rating of $a$ will then be a function of the values in this table, namely the degrees of utility obtained so far and the similarities

$$\sigma_\imath(a) = \sigma_{\mathcal{Q}\times\mathcal{A}}((p_0, a), (p_\imath, a_\imath))$$

between the already encountered problem–act tuples and the new tuple $(p_0, a)$. The criterion (7.4), for instance, is given by the weighted sum

$$V(a) = \sum_{\imath=1}^{n} \sigma_\imath(a) x_\imath.$$

Clearly, this criterion and table (7.12) remind one of expected utility theory. In fact, the context $x$ plays formally the role of the probability distribution on the set of states of nature, and the degrees of similarity $\sigma_\imath$ correspond to degrees of utility in EUT.

For the NN-rules (7.8) and (7.10) we can show representation theorems similar to the one obtained in [169]. Consider the following axioms, which are basically formulated in terms of *contexts*[10] ($\succ_x$ and $\simeq_x$ denote the asymmetric and symmetric part of $\succeq_x$, respectively):

A1 Order: $\succeq_x$ is complete and transitive for all $x \in \mathfrak{R}^n$.

A2 Continuity: For all $(x^k)_{k\geq 1} \subset \mathfrak{R}^n$ and all $a, b \in \mathcal{A}$ it holds true that

$$\left(x^k \to x \;\wedge\; \forall\, k \geq 1 \,:\, a \succeq_{x^k} b\right) \quad \Rightarrow \quad a \succeq_x b.$$

A3 Additivity: For all $x, y \in \mathfrak{R}^n$ and $a, b \in \mathcal{A}$ it holds true that

$$a \succ_x b \wedge a \succeq_y b \;\Rightarrow\; a \succ_{x+y} b.$$

A4 Neutrality: For all $a, b \in \mathcal{A}$ it holds true that $a \simeq_{(0,\ldots,0)} b$.

A5 Diversity: For all distinct acts $a, b, c, d \in \mathcal{A}$ a vector $x \in \mathfrak{R}^n$ exists such that

$$a \succ_x b \succ_x c \succ_x d.$$

The following result has been shown in [169]: A1–A5 imply the existence of vectors $\omega(a) = (\omega_1(a), \ldots, \omega_n(a))$ for all $a \in \mathcal{A}$ such that

$$a \succeq_x b \;\Leftrightarrow\; \sum_{\imath=1}^{n} \omega_\imath(a) \cdot x_\imath \geq \sum_{\imath=1}^{n} \omega_\imath(b) \cdot x_\imath, \tag{7.13}$$

where the $x_\imath$ are the utility degrees in (7.12). Moreover, the vectors $\omega(a)$ are unique up to an affine transformation. Of course, the weights $\omega_\imath(a)$ can be interpreted as the similarity degrees $\sigma_\imath(a)$ in (7.12).

[10] This contrasts with classical decision-theoretic models which are formalized in terms of acts (in the SAVAGE setting) or probabilistic lotteries (in the VON NEUMANN-MORGENSTERN framework).

The valuations (7.8) and (7.10) are obviously special cases of the weighted sum in (7.13). In order to obtain a set of axioms which imply a nearest neighbor representation it is hence possible to extend A1–A5 in such a way that some of the weights $\omega_\imath$ become 0. Consider the following axiom (the $k$-th entry of the vector $e_k$ is 1 and all other entries are 0):

A6 For all acts $a, b, c \in \mathcal{A}$, $x \in \Re^n$, $\gamma \geq 0$ and $1 \leq k \leq n$ it holds true that

$$c \succ_x a \;\wedge\; c \succ_x b \;\Rightarrow\; c \succ_{x+\gamma e_k} a \;\vee\; c \succ_{x+\gamma e_k} b. \tag{7.14}$$

In a certain sense, the meaning of A6 is opposite to that of axiom A5. The latter demands that a set of acts can be put in any order by defining the context appropriately. As opposed to this, A6 demands that a certain modification of the context, namely the increase of one utility degree $x_k$, can only have a limited influence: It can reverse but one of the preferences in the antecedent part of implication (7.14).

**Lemma 7.1.** Suppose A1-A6 to hold and let $1 \leq k \leq n$. The vector

$$\lambda = (\lambda_1, \ldots, \lambda_m) = (\omega_k(a_1), \ldots, \omega_k(a_m)),$$

where $m = \operatorname{card}(\mathcal{A}) \geq 4$, is of the form

$$\lambda = \alpha e_{\imath_0} + \beta \tag{7.15}$$

for some $1 \leq \imath_0 \leq m$, $\alpha \geq 0$ and $\beta \in \Re$. □

**Proof.** Consider a permutation $\pi$ of $\{1, \ldots, m\}$ such that

$$\lambda_{\pi(1)} \geq \lambda_{\pi(2)} \geq \ldots \geq \lambda_{\pi(m)}. \tag{7.16}$$

We obviously have $\lambda_{\pi(2)} = \ldots = \lambda_{\pi(m)}$ if (7.15) holds. Suppose by way of negation that

$$\lambda_{\pi(2)} \geq \ldots \geq \lambda_{\pi(\jmath-1)} > \lambda_{\pi(\jmath)} \geq \ldots \geq \lambda_{\pi(m)}.$$

Axiom A5 guarantees the existence of $x \in \Re^n$ such that $a_{\pi(\jmath)} \succ_x a_{\pi(1)}$ and $a_{\pi(\jmath)} \succ_x a_{\pi(2)}$. Since $\omega_k(a_{\pi(\jmath)}) = \lambda_{\pi(\jmath)} < \lambda_{\pi(1)} = \omega_k(a_{\pi(1)})$ and $\omega_k(a_{\pi(\jmath)}) = \lambda_{\pi(\jmath)} < \lambda_{\pi(2)} = \omega_k(a_{\pi(2)})$, there is obviously some $\gamma > 0$ such that

$$\sum_{\imath=1}^{n} \omega_\imath(a_{\pi(\imath_0)}) \cdot (x_\imath + \gamma e_k) > \sum_{\imath=1}^{n} \omega_\imath(a_{\pi(\jmath)}) \cdot (x_\imath + \gamma e_k)$$

for $\imath_0 = 1$ and $\imath_0 = 2$. This means $a_{\pi(1)} \succ_{x+\gamma e_k} a_{\pi(\jmath)}$ and $a_{\pi(2)} \succ_{x+\gamma e_k} a_{\pi(\jmath)}$ according to (7.13) and, hence, contradicts A6. Consequently, the representation (7.15) must hold with $\imath_0 = \pi(1)$, $\alpha = \lambda_{\pi(1)} - \lambda_{\pi(2)}$ and $\beta = \lambda_{\pi(2)}$. □

**Theorem 7.2.** Consider a decision problem with card($\mathcal{A}$) $\geq 4$. The preference relations $\succeq_x$ can be represented by (7.10) iff they satisfy A1–A6. □

**Proof.** It is readily verified that the preference relations $\succeq_x$ defined by (7.10) satisfy A1–A6. Concerning the converse direction, we make use of Lemma 7.1 and the fact that $\beta$ in (7.15) can be set to 0 without loss of generality. In fact, the variation of $\beta$ does not influence the relation on the right-hand side of (7.13). Thus, for each $1 \leq k \leq n$ there is at most one $1 \leq \imath_0 \leq m$ such that $\omega_k(a_{\imath_0}) \neq 0$. We hence obtain the representation (7.10) by letting

$$\mathsf{NN}_{p_0,\mathcal{A}}(p_\imath, a_\imath) = \{a \in \mathcal{A} \,|\, \omega_\imath(a) > 0\}$$

for all $(p_\imath, a_\imath) \in \mathcal{M}^\downarrow$. □

Note that a value $\omega_\imath(a) > 0$ is interpreted as the similarity between $(p_\imath, a_\imath)$ and $(p_0, a) = \mathsf{NN}_{p_0,\mathcal{A}}(p_\imath, a_\imath)$. It hence corresponds to the value $\sigma_\imath(a)$ in (7.12). It is clear, however, that the complete similarity relation $\sigma_{\mathcal{Q}\times\mathcal{A}}$ cannot be determined by the preferences $\succeq_x$. In fact, $\omega_\imath(b) = 0$ does not necessarily mean that $\sigma_{\mathcal{Q}\times\mathcal{A}}((p_\imath, a_\imath), (p_0, b)) = 0$ but only implies $\sigma_{\mathcal{Q}\times\mathcal{A}}((p_\imath, a_\imath), (p_0, b)) < \omega_\imath(a)$. This is caused by the behavior of a decision maker applying the NN principle. According to (7.9) it concentrates on the nearest neighbors of the observations but completely ignores other acts to which it assigns a relevance of 0. Thus, the preferences $\succeq_x$ can determine only the *relevance* of a case but not its *similarity* to $(p_0, a_0)$.

Now, consider again the decision rule (7.8). Axiom A5 is obviously not satisfied in connection with this criterion. Indeed, we have

$$b \succeq_x c \succeq_x d \quad \text{or} \quad d \succeq_x c \succeq_x b$$

for all $x \in \mathfrak{R}^n$ if the acts $b, c, d \in \mathcal{A}$ have the same nearest neighbor $(p, a)$ and if

$$\sigma_{\mathcal{Q}\times\mathcal{A}}((p, a), (p_0, b)) < \sigma_{\mathcal{Q}\times\mathcal{A}}((p, a), (p_0, c)) < \sigma_{\mathcal{Q}\times\mathcal{A}}((p, a), (p_0, d)).$$

Observe, however, that the act $c$ will then be ignored by the decision maker in the sense that it is not chosen anyway (except perhaps if $V(b) = V(c) = V(d) = 0$). Besides, act $b$ becomes interesting only if all acts have a negative (estimated) utility according to (7.8), a situation that can formally be avoided (see below). We can hence restrict the decision rule (7.8) to a set $\mathcal{A}_{p_0}$ of acts as follows: For the problem–act tuple $(p_\imath, a_\imath) \in \mathcal{M}^\downarrow$ define

$$A_{p_0}(p_\imath, a_\imath) = \arg \max_{a \in \mathcal{A} \,:\, \mathsf{NN}_{\mathcal{M}}(p_0, a) = (p_\imath, a_\imath)} \sigma_{\mathcal{Q}\times\mathcal{A}}((p_0, a), (p_\imath, a_\imath))$$

whenever the set on the right-hand side is not empty. For the sake of simplicity, we again assume $A_{p_0}(p_\imath, a_\imath)$ to be unique. The set $\mathcal{A}_{p_0}$ is then defined as

$$\mathcal{A}_{p_0} = \{A_{p_0}(p_\imath, a_\imath) \,|\, 1 \leq \imath \leq n, A_{p_0}(p_\imath, a_\imath) \text{ exists}\}. \tag{7.17}$$

It can be thought of as the set of *relevant* acts. As already suggested above, a decision based on (7.8) might appear somewhat peculiar if $V(a_0) < 0$ for all $a_0 \in \mathcal{A}$. Observe, however, that this problem can formally be avoided by means of a proper definition of acts. For instance, one might introduce a new act $a^*$ which stands for "doing anything" or "trying something completely new." When adding a dummy case $(p^*, a^*, 0)$ to $\mathcal{M}$, $V(a^*) = 0$ is guaranteed by letting $\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p^*,a^*)) = 1$ if $(p,a) = (p^*,a^*)$ and 0 otherwise. Thus, $a^*$ is preferred to each act with a negative estimated utility. This is clearly in accordance with the idea of an aspiration level in [167].

Note that each act $a \in \mathcal{A}_{p_0}$ in (7.17) has a unique nearest neighbor in $\mathcal{M}^{\downarrow}$. Moreover, for each $(p_\imath, a_\imath) \in \mathcal{M}^{\downarrow}$ there is at most one act $a \in \mathcal{A}$ such that $(p_\imath, a_\imath)$ is the nearest neighbor of $(p_0, a)$ (which implies $\text{card}(\mathcal{A}_{p_0}) \leq \text{card}(\mathcal{M})$). It is hence obvious that A5 is satisfied for $\mathcal{A}_{p_0}$. Besides, it is not difficult to show that the preference relations induced by (7.8) also satisfy the following axiom:

A7 For all $x, y \in \mathfrak{R}^n$ and $a, b \in \mathcal{A}$ it holds true that

$$a \succeq_x b \wedge a \succeq_y b \;\Rightarrow\; a \succeq_{\max\{x,y\}} b,$$

where the maximum of the vectors $x$ and $y$ is defined component-wise.

**Theorem 7.3.** Consider a decision problem with $\text{card}(\mathcal{A}) \geq 4$. The preference relations $\succeq_x$ can be represented by (7.8) iff they satisfy A1–A7. □

**Proof.** Again, A1–A7 are obviously satisfied when representing $\succeq_x$ by (7.8). In order to show the converse direction suppose A1–A7 to be satisfied. Given A1–A6, is has been shown in Theorem 7.2 that (7.13) holds in such a way that $\omega_\imath(a)\omega_\imath(b) = 0$ for all acts $a \neq b$. In order to establish a representation of $\succeq_x$ by (7.8), we further have to show that $\imath \neq \jmath \Rightarrow \omega_\imath(a)\omega_\jmath(a) = 0$ for all acts $a$. Thus, assume the existence of an act $a$ such that $\omega_\imath(a) > 0$ and $\omega_\jmath(a) > 0$ for $1 \leq \imath \neq \jmath \leq n$. Moreover, let the contexts $x$ and $y$ be defined as follows:

$$x_k = \begin{cases} \omega_\jmath(a) & \text{if} \quad k = \imath \\ -\omega_\imath(a) & \text{if} \quad k = \jmath \\ 0 & \text{if} \quad \imath \neq k \neq \jmath \end{cases}, \quad y_k = \begin{cases} -\omega_\jmath(a) & \text{if} \quad k = \imath \\ \omega_\imath(a) & \text{if} \quad k = \jmath \\ 0 & \text{if} \quad \imath \neq k \neq \jmath \end{cases}.$$

It is readily verified that $V(a) = 0$ in both contexts. Moreover, $V(b) = 0$ does also hold true for all other acts since $b \neq a$ entails $\omega_\imath(b) = \omega_\jmath(b) = 0$. Thus, $b \succeq_x a$ and $b \succeq_y a$ for any act $b \neq a$. In the context $\max\{x,y\}$, however, we have $V(a) = 2\omega_\imath(a)\omega_\jmath(a) > 0$ and, hence, $a \succ_{\max\{x,y\}} b$. This contradicts A7. □

## 7.3 Fuzzy modeling of case-based decisions

Case-based decision making has been realized in [101] as a kind of case-based approximate reasoning. This approach is in line with methods of qualitative decision theory. In fact, the assumption that uncertainty and preference can be quantified by means of, respectively, a precise probability measure and a cardinal utility function (as it is assumed in classical decision theory) does often appear unrealistic. As opposed to (7.2), the approach discussed in this section only assumes an ordinal setting for modeling decision problems, i.e., ordinal scales for assessing preference and similarity. This interpretation should be kept in mind, especially since both scales will subsequently be taken as (subsets of) the unit interval.

### 7.3.1 Basic measures for act evaluation

Let $\rightsquigarrow$ be a multiple-valued implication connective. Given a memory $\mathcal{M}$ and a new problem $p_0$, the following (estimated) utility value is assigned to an act $a \in \mathcal{A}$:

$$V^{\downarrow}_{p_0,\mathcal{M}}(a) \overset{\mathrm{df}}{=} \min_{(p,a,r)\in\mathcal{M}} \sigma_{\mathcal{Q}}(p,p_0) \rightsquigarrow u(r). \tag{7.18}$$

This valuation supports the idea of finding an act $a$ which has *always* resulted in good outcomes for problems similar to the current problem $p_0$. Indeed, (7.18) can be considered as a generalized truth degree of the claim that "whenever $a$ has been applied to a problem $p$ similar to $p_0$, the corresponding outcome has yield a high utility." An essential idea behind (7.18) is that of avoiding the accumulation and compensation effect caused by the decision criterion (7.2) (cf. Section 7.1),[11] since these effects do not always seem appropriate (cf. Section 7.6).

As a special realization of (7.18) the valuation

$$V^{\downarrow}_{p_0,\mathcal{M}}(a) \overset{\mathrm{df}}{=} \min_{(p,a,r)\in\mathcal{M}} \max\{n(h(\sigma_{\mathcal{Q}}(p,p_0))), u(r)\},$$

is proposed, where $h$ is an order-preserving function which assures the linear scales of similarity and preference to be commensurable and $n$ is the order-reversing function of the similarity scale. By taking $n$ as $x \mapsto 1-x$ in $[0,1]$ and $h$ as the identity, we obtain

$$V^{\downarrow}_{p_0,\mathcal{M}}(a) \overset{\mathrm{df}}{=} \min_{(p,a,r)\in\mathcal{M}} \max\{1-\sigma_{\mathcal{Q}}(p,p_0), u(r)\}. \tag{7.19}$$

This criterion can obviously be seen as a qualitative counterpart to (7.2). Besides, the criterion

$$V^{\uparrow}_{p_0,\mathcal{M}}(a) \overset{\mathrm{df}}{=} \max_{(p,a,r)\in\mathcal{M}} \min\{\sigma_{\mathcal{Q}}(p,p_0), u(r)\} \tag{7.20}$$

is introduced as an *optimistic* counterpart to (7.19). It can be seen as a formalization of the idea to find an act $a$ for which there is at least one problem which is similar to $p_0$ and for which $a$ has led to a good result. Again, let us mention that expressions (7.19) and (7.20) are closely related to decision criteria which

[11] Note that the accumulation effect is also the main motivation for the normalization (7.3).

have recently been derived in [123] in connection with an axiomatic approach to qualitative decision making under uncertainty.

In the more general context of problem–act similarity, the decision rules (7.19) and (7.20) become

$$V^{\downarrow}_{p_0,\mathcal{M}}(a_0) \overset{\text{df}}{=} \min_{(p,a,r)\in\mathcal{M}} \max\{1-\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)),u(r)\}, \tag{7.21}$$

$$V^{\uparrow}_{p_0,\mathcal{M}}(a_0) \overset{\text{df}}{=} \max_{(p,a,r)\in\mathcal{M}} \min\{\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)),u(r)\}. \tag{7.22}$$

In order to make the basic principles underlying the above criteria especially obvious, suppose the qualitative utility scale to be given by $U=\{0,1\}$. That is, only a crude distinction between "bad" and "good" outcomes is made. (7.21) and (7.22) can then be simplified as follows:

$$V^{\downarrow}_{p_0,\mathcal{M}}(a_0) \overset{\text{df}}{=} 1-\max_{(p,a,r)\in\mathcal{M}\,:\,u(r)=0} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)), \tag{7.23}$$

$$V^{\uparrow}_{p_0,\mathcal{M}}(a_0) \overset{\text{df}}{=} \max_{(p,a,r)\in\mathcal{M}\,:\,u(r)=1} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)). \tag{7.24}$$

According to (7.23), the decision maker only takes cases $(p,a,r)$ with bad outcomes into account. An act $a_0$ is discounted whenever $(p_0,a_0)$ is similar to a corresponding problem–act tuple $(p,a)$. Thus, the agent is cautious and looks for an act that it does not associate with a bad experience. According to (7.24), it only considers the cases with good outcomes. An act $a_0$ appears promising if $(p_0,a_0)$ is similar to a tuple $(p,a)$ which has yielded a good result. In other words, the decision maker is more adventurous and looks for an act that it associates with a good experience.

### 7.3.2 Modification of the basic measures

As noted in [125], (7.19) makes only sense if the memory contains at least one problem $p$ such that $\sigma_{\mathcal{Q}}(p,p_0)=1$ and $a$ has been chosen for solving $p$. Otherwise, it may happen that (7.19) is very high even though none of the problems contained in the memory is similar to the current problem $p_0$.[12] Particularly,

$$\left(\left\{p\in\mathcal{Q}\,|\,(p,a,r)\in\mathcal{M}\;\wedge\;\sigma_{\mathcal{Q}}(p,p_0)>0\right\}=\emptyset\right) \;\Rightarrow\; \left(V^{\downarrow}_{p_0,\mathcal{M}}(a)=1\right),$$

which does not seem satisfactory.

Modifications of (7.19) and its optimistic counterpart have been proposed in order to cope with these difficulties. The modified version of (7.19) is based on some kind of *normalization* of the similarity function for each act $a$ and a discounting

[12] Notice that the averaged similarity criterion (7.3) suffers from the same drawback.

which takes the absence of problems similar to $p_0$ into account. More precisely, the modified measure is given by

$$V^{\downarrow}_{p_0,\mathcal{M}}(a) = \min\left\{ h_{\mathcal{M}}(a,p_0), \min_{(p,a,r)\in\mathcal{M}} \max\{1-\sigma^a_{\mathcal{Q}}(p,p_0), u(r)\}\right\}, \qquad (7.25)$$

where

$$h_{\mathcal{M}}(a,p_0) = \max_{(p,a,r)\in\mathcal{M}} \sigma_{\mathcal{Q}}(p,p_0),$$

and $\sigma^a_{\mathcal{Q}}(\cdot,p_0)$ denotes a normalization[13] of $\sigma_{\mathcal{Q}}(\cdot,p_0)$, e.g.,

$$\sigma^a_{\mathcal{Q}}(p,p_0) = \begin{cases} 1 & \text{if} \quad \sigma_{\mathcal{Q}}(p,p_0) = h_{\mathcal{M}}(a,p_0) \\ \sigma_{\mathcal{Q}}(p,p_0) & \text{if} \quad \sigma_{\mathcal{Q}}(p,p_0) < h_{\mathcal{M}}(a,p_0) \end{cases}.$$

The idea behind (7.25) is that the willingness of a decision maker to choose act $a$ is upper-bounded by the existence of problems which are completely similar to $p_0$ and to which $a$ has been applied. Moreover, $\sigma_{\mathcal{Q}}(\cdot,p_0)$ is normalized in order to obtain a meaningful degree of inclusion. Thus, (7.25) corresponds to the compound condition that there are problems similar to $p_0$ to which act $a$ has been applied and the problems which are most similar to $p_0$ are among the problems for which $a$ has led to good results. Observe that (7.19) is retrieved from (7.25) as soon as $h_{\mathcal{M}}(a,p_0) = 1$. Moreover, note that a corresponding modification can also be defined for (7.20):

$$V^{\uparrow}_{p_0,\mathcal{M}}(a) = \max\left\{ 1 - h_{\mathcal{M}}(a,p_0), \max_{(p,a,r)\in\mathcal{M}} \min\{\sigma^a_{\mathcal{Q}}(p,p_0), u(r)\}\right\}. \qquad (7.26)$$

The criteria (7.25) and (7.26) guarantee that $V^{\downarrow}_{p_0,\mathcal{M}}(a) \leq V^{\uparrow}_{p_0,\mathcal{M}}(a)$ which is not necessarily the case for (7.19) and (7.20).

### 7.3.3 Interpretation of the decision criteria

As opposed to (7.2), the criteria (7.19) and (7.20) do obviously not focus on some kind of average performance, which hardly makes sense within an ordinal setting. Rather, they should be considered from the same point of view as qualitative decision rules such as Maximin [123]. Indeed, the application of (7.18) seems reasonable, for instance, if an agent aims at minimizing the occurrence of worst case outcomes in competition with other agents or if only an ordinal preference relation on outcomes is assumed [52].

We shall now propose two interpretations of (7.19).[14] The first one is that of an approximation of a (generalized) Maximin evaluation: Observe that we can write (7.19) as

[13] Note that this normalization is again defined for each act individually.

[14] These interpretations can be transferred to (7.20) in a straightforward way.

$$V^{\downarrow}_{p_0,\mathcal{M}}(a) = \min_{0 \leq k \leq m} \max\{1 - \sigma_k, v_k\}, \tag{7.27}$$

where the values $0 = \sigma_0 < \sigma_1 < \ldots < \sigma_m = 1$ constitute the (finite) set $\{\sigma_{\mathcal{Q}}(p,p') \,|\, p,p' \in \mathcal{Q}\}$ of possible similarity degrees of problems and

$$v_k = \min V_k = \min\{u(r) \,|\, (p,a,r) \in \mathcal{M}, \sigma_{\mathcal{Q}}(p,p_0) = \sigma_k\}$$

is the lowest utility obtained in connection with act $a$ for problems which are $\sigma_k$-similar to $p_0$. Moreover, $v_k = 1$ by definition if $V_k = \emptyset$ (which is just the reason for the problem that (7.19) becomes large if no similar observations have been made).

According to (7.27), the valuation (7.19) of an act is completely determined by the lower bounds $v_k$ ($0 \leq k \leq m$) which are derived from the memory $\mathcal{M}$ (and discounted according to respective degrees of similarity). This reveals that (7.19) can indeed be seen as some kind of "experience-based" approximation of the Maximin principle. The case in which all problems are completely similar makes this especially apparent. Then, (7.19) evaluates an act $a$ simply according to the worst consequence observed so far. More generally, the value $v_k$ can be seen as an estimation of the lower utility bound

$$w_k = \min\{u(r(p,a)) \,|\, p_0 \neq p \in \mathcal{Q}, \sigma_{\mathcal{Q}}(p,p_0) = \sigma_k\},$$

i.e., the smallest degree of utility which can be obtained in connection with act $a$ for (not necessarily encountered) problems from $\mathcal{Q}$ which are $\sigma_k$-similar to $p_0$. Then, $V^{\downarrow}_{p_0,\mathcal{M}}(a)$ can be interpreted as an approximation of

$$W^{\downarrow}_{p_0}(a) = \min_{0 \leq k \leq m} \max\{1 - \sigma_k, w_k\},$$

which defines a case-based generalization of a Maximin-evaluation. In fact, $W^{\downarrow}_{p_0}(a)$ is equal to $V^{\downarrow}_{p_0,\mathcal{M}}(a)$ if $a$ has already been applied to all problems (up to $p_0$) from $\mathcal{Q}$, i.e., if $\{p \,|\, \exists\, r \in \mathcal{R} : (p,a,r) \in \mathcal{M}\} = \mathcal{Q} \setminus \{p_0\}$.

According to a second (more logic-oriented) interpretation, (7.18) might be seen as the (generalized) truth degree of a proposition characterizing the decision maker's preferences concerning acts. In our case, those acts are preferred which have always resulted in good outcomes for similar problems. Then, (7.18) defines the degree to which an act meets the requirements and, hence, induces a corresponding preference relation over acts. In a certain sense, this approach can be seen as a "compiled" decision model which skips the estimation of utility and relates similarity or, more generally, certain properties of an act to preference more directly. That is to say, the agent already knows which properties a preferred act should have. The idea of such a compiled model becomes even more obvious if we consider (crisp) rules of the form "if the problem has property $x$ then choose an act with property $y$", such as "if it looks rainy then take an umbrella with you." Rules of this kind are often set up if a decision problem is solved frequently. They represent a sort of routine decision and reflect the agent's knowledge that a decision analysis, i.e., the estimation of utility degrees for all decisions, would result in choosing a certain act if the problem has a related property anyway.

Even though formally equivalent, the two interpretations are different from a semantical point of view. For instance, interpreting the value $V(a)$ assigned to an act $a$ which has not yet been tried as a degree to which this act meets the agent's idea of an "ideal" decision seems less critical than viewing this value as an estimated utility. In fact, the latter is merely a "default utility." As opposed to this, the former interpretation does principally not assume observations at all. Rather, $V(a)$ can be seen as reflecting the agent's attitude toward uncertainty. Assigning a high default value to $a$ then simply means that a not yet applied act seems attractive and, hence, amounts to model an uncertainty-prone decision maker who is willing to try new acts.

## 7.4 Fuzzy quantification in act evaluation

In some situations, the extremely pessimistic and optimistic nature of the criteria (7.19) and (7.20), respectively, might appear at least as questionable as the accumulation in (7.2). Here we shall propose a generalization of the decision rule (7.19) which is a weakening of the demand that an act has *always* produced good results for similar problems. In fact, one might already be satisfied if $a$ turned out to be a good choice *for most* similar problems, thus allowing for a few exceptions [125]. In other words, the idea is to relax the universal "for all" quantifier. Observe that a similar generalization of (7.20), which replaces "there exists" by "there are at least several" and, hence, corresponds to a strengthening of this decision principle, seems reasonable as well. It can be obtained analogously.

Consider a finite set $A$ of cardinality $m = |A|$. In connection with propositions of the form "most elements of $A$ have property $X$" the fuzzy quantifier "most" can be formalized by means of a fuzzy set [132, 403],[15] the membership function $\mu : \{0, 1, \ldots, m\} \longrightarrow [0, 1]$ of which satisfies

$$\forall\, 1 \leq k \leq m-1 \,:\, \mu(k) \leq \mu(k+1) \qquad \text{and} \qquad \mu(m) = 1. \tag{7.28}$$

The special case "for all" then corresponds to $\mu(k) = 0$ for $0 \leq k \leq m-1$ and $\mu(m) = 1$. Given some $\mu$ satisfying (7.28), we define an associated membership function $\overline{\mu}$ by $\overline{\mu}(0) = 0$ and $\overline{\mu}(k) = 1 - \mu(k-1)$ for $1 \leq k \leq m$ (see e.g. [109]). A membership degree $\overline{\mu}(k)$ can then be interpreted as quantifying the importance that the property $X$ is satisfied for $k$ (out of the $m$) elements.

Consider a memory $\mathcal{M}$ of cases, a problem $p_0 \in \mathcal{Q}$, an act $a \in \mathcal{A}$, and let $\mathcal{M}_a = \{(p', a', r') \in \mathcal{M} \,|\, a = a'\}$. Moreover, let $\mu$ formalize the above-mentioned "for most" concept. A reasonable generalization of (7.19) is then given by

[15] Other possibilities of expressing a fuzzy quantifier exist as well, including the use of order-statistics [300] and an ordered weighted minimum or maximum [135].

$$V_{p_0,\mathcal{M}}(a) = \min_{0 \leq k \leq |\mathcal{M}_a|} \max \left\{ 1 - \overline{\mu}(k), \delta_a(k) \right\}, \tag{7.29}$$

where

$$\delta_a(k) = \max_{\mathcal{M}' \subset \mathcal{M}_a \,:\, |\mathcal{M}'| = k} \; \min_{(p,a,r) \in \mathcal{M}'} \max\{1 - \sigma_{\mathcal{Q}}(p, p_0), u(r)\}$$

defines the degree to which "the act $a$ has induced good outcomes for similar problems $k$ times." The extent to which a (small) degree $\delta_a(k)$ decreases the overall valuation of $a$ is upper bounded by $1 - \overline{\mu}(k)$, i.e., by the respective level of (un-)importance. Observe that we do not have to consider all subsets $\mathcal{M}' \subset \mathcal{M}_a$ of size $k$ for deriving $\delta_a(k)$. In fact, for computing $V_{p_0,\mathcal{M}}(a)$ it is reasonable to arrange the $m = |\mathcal{M}_a|$ values $v = \max\{1 - \sigma_{\mathcal{Q}}(p, p_0), u(r)\}$ in a non-increasing order $v_1 \geq v_2 \geq \ldots \geq v_m$. Then, (7.29) is equivalent to

$$V_{p_0,\mathcal{M}}(a) = \min_{0 \leq k \leq |\mathcal{M}_a|} \max \left\{ 1 - \overline{\mu}(k), v_k \right\},$$

where $v_0 = 1$.

The generalized criterion (7.29) can be useful, e.g., in connection with the idea of repeated decision making which arises quite naturally in connection with a case-based approach to decision making. We might think of different scenarios in which repeated problem solving becomes relevant. A simple model emerges from the assumption that problems are chosen repeatedly from $\mathcal{Q}$ according to some selection process which is not under the control of the agent, such as the repeated (and independent) selection of problems according to some probability measure. More generally, the problem faced next by the agent might depend on the current problem and the act which is chosen for solving it. A Markov Decision Process extended by a similarity measure over states (which correspond to problems) may serve as an example. Besides, we might consider case-based decision making as a reasonable strategy within a (repeated) game playing framework like the iterated prisoner's dilemma [19].

As a concrete example let us consider a very simple model of repeated decision making: Suppose that the agent faces the same problem $p$ repeatedly and that the result associated with an act $a \in \mathcal{A} = \{a_1, a_2, a_3\}$ depends on a state of nature $\omega \in \Omega = \{\omega_1, \omega_2, \omega_3\}$. The state $\omega$ is assumed to be chosen randomly (every time) and is not part of the problem description. We assume the probability for $\omega = \omega_3$, which is also not known to the decision maker, to be positive but relatively small. Moreover, the results (= utilities) associated with act–state tuples shall be specified as follows:

| | $\omega_1$ | $\omega_2$ | $\omega_3$ |
|---|---|---|---|
| $a_1$ | 1 | 1 | 0 |
| $a_2$ | 1 | 0 | 0 |
| $a_3$ | 0 | 0 | 0 |

Recall that 0 and 1 are interpreted as ordinal degrees of utility; they only indicate that one outcome is preferred to the other one, which might be encoded by $-1$ and 1 as well.[16]

Now, since $a_1$ dominates $a_2$ and $a_3$ (strictly), it is obviously the best choice. Observe, however, that the valuation of an act $a$ according to (7.19) simply corresponds to the worst outcome observed in connection with this act, i.e.

$$V^{\downarrow}_{p_0,\mathcal{M}}(a) = \begin{cases} 0 & \text{if} \quad (p,a,0) \in \mathcal{M} \\ 1 & \text{if} \quad (p,a,0) \not\in \mathcal{M} \end{cases} .$$

Thus, we have $V^{\downarrow}_{p_0,\mathcal{M}}(a_1) = 0$ as soon as $a_1$ has been selected for solving $p$ and $\omega = \omega_3$. From this moment of time, $a_1$ and, sooner or later, $a_2$ and $a_3$ are rated equally and an act might be selected, e.g., by flipping a coin. In other words, the problem which occurs when basing decisions on (7.19) is the fact that this criterion does not, in the long run, discriminate between two acts even though the first one strictly dominates the second one. It is interesting to compare this with the MAXIMIN rule which also does not discriminate between $a_1$ and $a_3$.[17] This, however, seems to be acceptable more easily than the same property for (7.19): If used in connection with one-shot decisions, the MAXIMIN rule does not memorize experience from previous problem solving epochs. As opposed to a case-based decision rule, it does not have the opportunity of learning and experimenting in the course of a repeated problem solving process.[18]

The aforementioned drawback can be avoided by (7.29) in conjunction with a proper formalization of the "for most" concept. In fact, since (7.29) allows for a few exceptions (and $\omega_3$ is assumed to occur but seldom) we will probably have $V_{p_0,\mathcal{M}}(a_2) = V_{p_0,\mathcal{M}}(a_3) = 0 < 1 = V_{p_0,\mathcal{M}}(a_1)$. Then, the relative frequency of selecting $a_1$ will converge toward 1 (instead of 1/3, as it would do in connection with a random choice between equally rated acts $a_1, a_2, a_3$). More precisely, suppose the "for all" quantifier to be defined such that it yields 1 if the property under consideration is satisfied in at least $100(1-\varepsilon)$ percent of the cases and 0 otherwise. In terms of our notation above, this means

$$\mu(k) = \begin{cases} 1 & \text{if} \quad k/m \geq 1-\varepsilon \\ 0 & \text{if} \quad k/m < 1-\varepsilon \end{cases} .$$

We will then have $V_{p_0,\mathcal{M}}(a_1) = 0$ if the proportion $\pi_m$ of cases in which $\omega_3$ has occurred in connection with $a_1$ exceeds $\varepsilon$, where $m$ is the number of times $a_1$ has been chosen. Otherwise, we have $V_{p_0,\mathcal{M}}(a_1) = 1$. The probability that $\pi_m > \varepsilon$ and, hence, the probability that $V_{p_0,\mathcal{M}}(a_1) = 0$ will be small if $\varepsilon$ is chosen sufficiently large in relation to the probability of the occurrence of $\omega_3$. On the other hand, $\varepsilon$

[16] This clearly exemplifies that the application of (7.2) does hardly make sense.

[17] A discrimination can be achieved by extensions of MAXIMIN, such as the ordinal decision rules DISCRIMIN and LEXIMIN [152].

[18] This argument is no longer valid in a game playing context. Then, however, MAXIMIN can be justified by the assumption of an opponent acting optimally.

should not be made too large since otherwise $V_{p_0,\mathcal{M}}(a_2) = 1$ as well, which means that $a_1$ and $a_2$ are rated equally. An interesting idea arising in this context, which leads to a further extension of the model, is that of *learning* an optimal "for most" concept (from a parameterized class of membership functions). This can be seen as the counterpart to learning an optimal aspiration level in CBDT [168]. In our example, where the membership function $\mu$ depends only on $\varepsilon$, this parameter itself can be considered as an aspiration level.

Notice that the probability of $\pi_m > \varepsilon$ decreases with $m$ if the probability that $\omega = \omega_3$ is smaller than $\varepsilon$. Thus, the probability of disqualifying $a_1$ is, if at all, relatively large at the beginning of a decision sequence, i.e., as long as $a_1$ has not been tried very often. This problem can be alleviated by means of a more flexible specification of the "for most" concept. Namely, the smaller the value of $m$, the less restrictive this concept should be specified in terms of the membership function $\mu$. The definition above, for instance, could be generalized such that $\varepsilon$ depends on $m$, i.e., $\mu(k) = 1$ if $k/m \geq \varepsilon_m$ and $\mu(k) = 0$ otherwise, with a non-increasing sequence $(\varepsilon_m)_{m \geq 0}$.

Let us now pass over from the (case-based) valuation of single acts (in the context of a certain problem) to the valuation of complete decision strategies. Of course, the question when to prefer a certain decision rule to an alternative criterion is by no means obvious in connection with the assumption of an ordinal setting for decision making. In fact, all kinds of "averaging" like, e.g., the derivation of the mean of the obtained utility values, are out of the question. Using the worst outcome, which might appear natural if (7.19) is seen as a kind of (case-based) analogue of the Maximin decision rule, seems critical as well. In fact, within a *case-based* decision framework it is principally not possible to fully realize the idea underlying this (pessimistic) principle. Namely, an agent knows the possible consequences of a decision only *after* having applied the corresponding act. Then, however, the worst outcome has already occurred. In other words, it is impossible for a case-based decision maker to avoid the worst outcome in any case or to choose acts according to a (proper) Maximin principle.

In connection with a model in which problems are chosen repeatedly according to some probability it seems reasonable to prefer a decision strategy $S$ to a strategy $S'$ if the former *dominates* the latter (stochastically) in the following sense: Let $U = \{u_1, u_2, \ldots, u_m\}$ such that $u_1 < u_2 < \ldots < u_m$ define the (linearly ordered) utility scale, and denote by $P_k^n(S)$ the probability of obtaining the utility $u_k$ in the $n$-th step of a decision sequence if strategy $S$ is used.[19] Then, $S$ dominates $S'$ (stochastically) if

$$\forall\, n \in \mathfrak{N}\; \forall\, 1 \leq k \leq m \; : \; \sum_{\imath=k}^{m} P_\imath^n(S') \leq \sum_{\imath=k}^{m} P_\imath^n(S). \tag{7.30}$$

[19] Observe that the sequences $(a(n))_{n \geq 1}$ of decisions and $(u(n))_{n \geq 1}$ of obtained outcomes resp. utility values are well-defined stochastic processes. In fact, for a (deterministic or stochastic) case-based decision procedure, the $n$-th decision is a function of the stochastic sequence of the first $n$ problems $(p(1), \ldots, p(n))$.

For our example above, we have $U = \{0, 1\}$, i.e., $P_0^n(S)$ and $P_1^n(S)$ simply correspond to the probability of obtaining a "bad" and a "good" outcome, respectively, in connection with the $n$-th decision. Moreover, a decision criterion $S$ is preferred to $S'$ in the sense of (7.30) if $P_1^n(S) \geq P_1^n(S')$ for all $n \in \mathfrak{N}$.

Appendix F shows simulation results for different decision strategies $S_\varepsilon$ which differ only with respect to the choice of $\varepsilon$, i.e., the definition of the "for most" quantifier. The states $\omega_1, \omega_2, \omega_3$ occur with probability 0.6, 0.3, and 0.1, respectively. Acts are evaluated according to (7.29), and ties between equally rated decisions are broken by coin flipping.

The results confirm the supposition that $\varepsilon$ should satisfy $0.1 < \varepsilon < 0.4$. The critical values are $\varepsilon = 0.1$ and $\varepsilon = 0.4$. For $\varepsilon < 0.1$, the agent is overly ambitious, and all acts will sooner or later be judged equally and, hence, $P_1^n(S_\varepsilon) \rightarrow 1/2$ as $n \rightarrow \infty$. Letting $0.4 < \varepsilon$ is "too tolerant" in the sense that $V_{p_0,\mathcal{M}}(a_2) = 1$ in the long term, which means that (7.29) does not differentiate between $a_1$ and $a_2$ and, therefore, $P_1^n(S_\varepsilon) \rightarrow 3/4$ as $n \rightarrow \infty$. Note that the estimation of $P_1^n(S_\varepsilon)$ from the sequence $(u(n))_{n \geq 1}$ of obtained utility values is a good starting point for learning an optimal value for $\varepsilon$, i.e., for choosing an optimal "for most" concept from $\{\mu_\varepsilon \,|\, 0 \leq \varepsilon \leq 1\}$.

## 7.5 A CBI framework of CBDM

CBDT as introduced in [167] is largely motivated by practical problems arising in connection with EUT, notably the considerable need of precise information for modeling decision problems. Indeed, the specification of an EUT model might often be complicated and expensive, especially when having to solve relatively novel decision problems. In this section, we shall propose a framework of CBDM which also makes use of case-based reasoning in order to alleviate this problem, but which remains closer to classical decision theory. Loosely speaking, the idea is to apply the methods of case-based inference (CBI) discussed in previous chapters in order to support the modeling of decision problems.

### 7.5.1 Generalized decision-theoretic setups

The basic EUT setup (in the finite case) can be illustrated in the form of a table as follows:

$$\begin{array}{c|cccc} & \rho_1 & \rho_2 & \dots & \rho_n \\ & \omega_1 & \omega_2 & \dots & \omega_n \\ \hline a_1 & u_{11} & u_{12} & \dots & u_{1n} \\ a_2 & u_{21} & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_m & u_{m1} & u_{m2} & \dots & u_{mn} \end{array} \tag{7.31}$$

The $\omega_j$ constitute the set $\Omega$ of states of nature, and each $\omega_j$ is assumed to occur with probability $\rho_j$. Choosing act $a_i$ yields utility $u_{ij}$ if the state of nature is $\omega_j$, which means that the expected utility of $a_i$ is given by $\sum_{j=1}^{n} \rho_j u_{ij}$. The expected utility framework can be generalized in order to deal with infinite sets of acts and/or states of nature. Subsequently, however, we assume $\mathcal{A}$ and $\Omega$ to be finite.

When modeling a decision problem, some of the information in (7.31) might be incomplete or even missing. This concerns mainly the probability distribution on $\Omega$ and the utility function $u : \mathcal{A} \times \Omega \longrightarrow U$ which assigns a utility degree to each tuple consisting of an act and a state of nature. The basic idea which is discussed in this section and which characterizes CBDM is the use of case-based inference for deriving corresponding estimations. Of course, this approach presupposes the existence of *cases*. As will be seen, there are different possibilities for defining a case, each of which leads to a different extension of the basic EUT setup.

For instance, let $\mathcal{Q}$ be a set of problems and suppose an EUT setup (7.31) to be associated with each problem $p \in \mathcal{Q}$:

$$\begin{array}{c|cccc}
 & \rho_1^p & \rho_2^p & \cdots & \rho_n^p \\
 & \omega_1 & \omega_2 & \cdots & \omega_n \\
\hline
a_1 & u_{11}^p & u_{12}^p & \cdots & u_{1n}^p \\
a_2 & u_{21}^p & u_{22}^p & \cdots & u_{2n}^p \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_m & u_{m1}^p & u_{m2}^p & \cdots & u_{mn}^p
\end{array} \tag{7.32}$$

A case is then defined as a triple $(p, \rho^p, u^p)$, where $\rho^p$ and $u^p$ denote the probability distribution and utility function associated with the problem $p$, respectively. (The set of acts, $\mathcal{A}$, and the set of states of nature, $\Omega$, are assumed to be fixed.) Within the framework of CBI, the problem $p$ corresponds to an *input*. Moreover, $\rho^p$ and $u^p$ mark the *outcome* associated with a case, which can hence be written as a tuple $\langle p, (\rho^p, u^p) \rangle$. Note that such a case reduces to a tuple of the form $\langle p, \rho^p \rangle$ or $\langle p, u^p \rangle$ if either $u^p$ or $\rho^p$ is fixed in advance.

Suppose a decision maker to have a memory $\mathcal{M}$ of cases at its disposal. Given a new problem $p_0$, it can then make use of case-based inference in order to support the specification of a related EUT setup. This approach relies on the CBI assumptions

– that similar problems give rise to similar probability distributions on $\Omega$, and/or

– that an act yields similar utilities for similar problems (under the same state of nature).

Example 7.4. Consider different types of urn experiments as an example: Let a state of nature, $\omega$, correspond to the number of black balls in a random sample of size $k$. The sample is drawn from an urn which contains a large number $K$ of balls, each of which is either black or white. An act $a$ corresponds to an estimation of $\omega$, and the utility $u(a, \omega)$ depends on the accuracy of this estimation, i.e., on the (absolute) difference $|a - \omega|$. Moreover, suppose that a problem is associated with the experimental conditions under which a sample is taken. The problems "simple selection with replacement" and "simple selection without replacement" can be considered as being similar if $k/K$ is small. Indeed, the hypergeometric distribution which defines $\rho$ in the latter case can then be approximated by the binomial distribution which is relevant if selected balls are replaced. From a CBI perspective, knowledge of the distribution $\rho$ for the first problem can hence be seen as valuable information for defining the EUT setup for the (somewhat more complicated) second problem. Observe that the utility function is assumed to be known (and identical) for both problems. □

An alternative approach is to consider a setting in which the probability over $\Omega$ and/or the utility function depend not only on the problem but also on the act:

$$\begin{array}{c|cccc} & \rho_1^{(p,a)} & \rho_2^{(p,a)} & \dots & \rho_n^{(p,a)} \\ & \omega_1 & \omega_2 & \dots & \omega_n \\ \hline a & u_1^{(p,a)} & u_2^{(p,a)} & \dots & u_n^{(p,a)} \end{array} \tag{7.33}$$

A case can then be seen as a tuple $\langle (p, a), \mu_{p,a} \rangle$, where $\mu_{p,a}$ is a probability distribution on $U$. This definition is in accordance with the idea of a non-deterministic CBI setup as introduced in Section 2.4.2, where a random outcome is associated with each input. It can be considered as a generalization of CBDT which assumes the outcome associated with a problem–act tuple $(p, a)$ to be deterministic. Thus, both frameworks (7.32) and (7.33) combine aspects from EUT and CBDT. The former, however, seems to be closer to EUT, whereas the latter is quite similar to CBDT.

Observe that a setup

$$\begin{array}{c|cccc} & \rho_1^{(p_0,a_0)} & \rho_2^{(p_0,a_0)} & \dots & \rho_n^{(p_0,a_0)} \\ \hline a_0 & u_1^{(p_0,a_0)} & u_2^{(p_0,a_0)} & \dots & u_n^{(p_0,a_0)} \end{array} \tag{7.34}$$

makes also sense within the original context of CBDT where a problem–act tuple has a unique outcome, i.e., if a case is a triple $(p, a, r)$ resp. a tuple $\langle (p, a), r \rangle$. Then, however, an unknown outcome (or utility) is not considered as a random variable (in the proper sense), and an uncertainty measure $\eta_{p_0,a_0}$ associated with a new problem $p_0$ and an act $a_0$ is interpreted as a quantification of a (subjective) belief concerning this outcome. Such a framework can be seen as defining an extended Bayesian approach in which CBI is used for assessing a (prior) measure of uncertainty over $\Omega$. Symbolically, it can be illustrated as follows:

$$\left.\begin{array}{c} (p_1, a_1, r_1), \dots, (p_n, a_n, r_n) \\ p_0, a_0 \\ \sigma_{\mathcal{Q} \times \mathcal{A}}, \sigma_{\mathcal{R}} \end{array}\right\} \quad \xrightarrow{\text{CBI}} \quad \eta_{p_0,a_0}. \tag{7.35}$$

### 7.5.2 Decision making using belief functions

The type of uncertainty measure derived in (7.35) depends on the way in which CBI is realized. Within the probabilistic framework of Section 4.5, for instance, the measure $\eta_{p_0,a_0}$ takes the form of a belief function:

$$\eta_{p_0,a_0} = \mathsf{Bel}(H, \mathcal{M}, (p_0, a_0)) = \sum_{\imath=1}^{n} \alpha_\imath \cdot \mathsf{Bel}_\imath(H, (p_0, a_0)),$$

where $H$ is a probabilistic similarity hypothesis and

$$\mathsf{Bel}_\imath(H, (p_0, a_0)) = \sigma_{\mathcal{R}}^{(-1)} \left(r_\imath, H(\sigma_{\mathcal{Q}\times\mathcal{A}}((p_0, a_0), (p_\imath, a_\imath)))\right)$$

denotes the belief function associated with the $\imath$-th case $(p_\imath, a_\imath, r_\imath) \in \mathcal{M}$. In this context, the probability distribution in (7.34) is replaced by a belief function. Consequently, the concept of an expected utility has to be generalized in order to evaluate an act. In other words, a framework of CBDM can be obtained by combining the CBI method of Section 4.5 and a generalization of expected utility based on belief functions. In recent years, several approaches to decision making on the basis of belief functions have been proposed in literature. Subsequently, we shall describe some of them very briefly.

Consider a belief function $\mathsf{Bel}$ on a set of outcomes, $\mathcal{R}$, and let $\mathsf{m}$ denote the mass distribution associated with $\mathsf{Bel}$. Moreover, let $\mathcal{F}$ be the set of focal elements of $\mathsf{m}$. A generalized expected utility can then be defined in terms of the Choquet integral

$$\int^{ch} u \, d\mathsf{Bel} = \int_0^{\infty} \mathsf{Bel}([u > t]) \, dt + \int_{-\infty}^{0} (\mathsf{Bel}([u > t]) - 1) \, dt, \tag{7.36}$$

where $[u > t] \overset{\text{df}}{=} \{r \in \mathcal{R} \,|\, u(r) > t\}$. This approach is a pessimistic strategy in the sense that (7.36) is equal to the minimum (the infimum in the non-finite case [390, 389]) of a class of associated classical expected utilities:

$$\int^{ch} u \, d\mathsf{Bel} = \min_{\mu \in \mathcal{P}_{\mathsf{Bel}}} \int u \, d\mu, \tag{7.37}$$

where

$$\mathcal{P}_{\mathsf{Bel}} = \{\mu \in \mathcal{P}(\mathcal{R}) \,|\, \forall\, X \subset \mathcal{R} \,:\, \mathsf{Bel}(X) \leq \mu(X)\} \tag{7.38}$$

is the set of probability measures over $\mathcal{R}$ compatible with $\mathsf{Bel}$. As (7.38) reveals, this approach favors a lower probability interpretation of belief functions.

Choquet expected utility now plays an important role in research on axiomatic non-expected utility. This research direction is motivated by the paradoxes of Allais [13] and Ellsberg [140] which call the validity of the assumptions underlying EUT into question. A "behavioral foundation" of Choquet expected utility in the context of decision making under uncertainty has first been given by SCHMEIDLER [334], who uses the decision-theoretic setup of ANSCOMBE & AUMANN [15]. A corresponding extension of the approach of SAVAGE [331] has been proposed in [166]. These works have been refined by several authors. An appealing axiomatic characterization of non-additive expected utility somehow unifying [334] and [166] has been developed in [330]. In [190] it is shown that a common characterizing property of this line of research is a certain weakening of SAVAGE's axioms which essentially restricts the well-known *sure thing principle* to so-called comonotonic acts.

Related models for decision making with belief functions have also been proposed in [210]. The axiomatic theory developed in [212] gives a foundation to these decision models. Here, situations are considered in which information is ambiguous and not fully probabilizable. It is argued that entirely vague information should be processed according to the (objective) symmetry principles of *complete ignorance* [16, 69] (rather than to the principle of insufficient reason). Again, the most important aspect of the decision-theoretic framework developed in [212] is a natural weakening of SAVAGE's sure thing principle [331]. It is shown that, within the resulting axiomatic setting, decisions can be represented by belief functions on outcomes. More precisely, a representation of a preference relation on the set of acts is of the form

$$f \mapsto \sum_{F \in \mathcal{F}} \mathsf{m}_f(F)\, v(r_F, R_F) \ , \tag{7.39}$$

where $\mathsf{m}_f$ is the Möbius transform (mass distribution) associated with the belief function induced by the act $f : \mathcal{A} \longrightarrow \mathcal{R}$ on the set of outcomes. Moreover, $r_F$ is the worst and $R_F$ is the best outcome within $F \in \mathcal{F}$. As as special case of (7.39) the functional

$$\sum_{F \in \mathcal{F}} \mathsf{m}_f(F)\ \left(\alpha(r_F, R_F)\, u(r_F) + (1 - \alpha(r_F, R_F))\, u(R_F)\right) \tag{7.40}$$

is proposed, where $u$ reflects the agent's attitude toward outcomes in decision under risk. The function $\alpha$ is interpreted as an index of the like or dislike of ambiguity.

A related generalization of the VON NEUMANN-MORGENSTERN framework has been proposed by JAFFRAY [211]. He combines the axioms of linear utility theory with axioms of rational decision making under *mixed uncertainty* [70] in order to justify a family of so-called Hurwicz $\alpha$-criteria. According to these criteria, a belief function $\mathsf{Bel}$ over the set of outcomes $\mathcal{R}$ is evaluated by

$$\alpha \inf\{\mathbb{E}_\mu(u) \mid \mu \in \mathcal{P}_{\mathsf{Bel}}\} + (1-\alpha) \sup\{\mathbb{E}_\mu(u) \mid \mu \in \mathcal{P}_{\mathsf{Bel}}\}, \tag{7.41}$$

where $\mathbb{E}_\mu(u)$ denotes the expected utility under the probability measure $\mu$. The use of Hurwicz criteria is also advocated by STRAT [361].

YAGER [405] defines a generalized expected utility of the form

$$\sum_{F \in \mathcal{F}} \mathsf{m}(F) \cdot \phi(F) \tag{7.42}$$

which makes use of a set-function $\phi : 2^{\mathcal{R}} \longrightarrow \mathfrak{R}$. The problem of assigning a degree of utility, $\phi(F)$, to a focal set $F$ is considered in the context of decision making under ignorance. It is proposed to solve this problem by applying an OWA (Ordered Weighted Average) operator[20] [404] to the collection $u(F)$ of utility degrees $u(r)$ $(r \in F)$.[21] That is, $\phi(F) = \text{OWA}(u(F))$. Special cases of this operator include the well-known decision rules

$$\begin{aligned}
\phi(F) &= \min u(F), \\
\phi(F) &= \alpha \min u(F) + (1-\alpha) \max u(F), \\
\phi(F) &= \sum_{r \in F} u(r) / \operatorname{card}(F).
\end{aligned}$$

Note that the set-function $\phi$ in (7.42) allows one to model the agent's decision behavior under complete ignorance in a more general way than the extreme (pessimistic) valuation by means of the Choquet integral (where always the worst case is assumed) or the Hurwicz criteria (7.40) (where $\phi(F)$ depends only on the worst and the best element in $F$).

As (7.37) shows, the use of Choquet integration comes down to deriving a classical expected utility based on the selection of a probability measure compatible with the belief function. In [350] it has been proposed to apply a generalization of LAPLACE's insufficient reason principle in order to select a corresponding distribution:

$$\mu(\{r\}) = \sum_{F \in \mathcal{F}} \mathbb{I}_F(r)\, m(F) / \operatorname{card}(F). \tag{7.43}$$

The transformation (7.43), which corresponds to the betting function (4.22) introduced in Section 4.5.1, has been justified axiomatically in the context of the *transferable belief model* which favors a purely subjective (and non-probabilistic) interpretation of belief functions. Note that (7.43) is the distribution of maximum entropy among $\mathcal{P}_{\mathsf{Bel}}$, i.e., it can also be derived from the principle of maximum entropy.

### 7.5.3 Possibilistic decision making

In Chapter 6, we have proposed a possibilistic method of case-based inference which makes use of implication-based fuzzy rules. According to this approach, uncertainty concerning the outcome $r_0$ is characterized by means of a possibility distribution:

[20] Operators of this type are also known as linear order statistics in the field of robust statistics.

[21] Each outcome $r \in F$ contributes exactly one element to $u(F)$, i.e., the same utility degree might appear several times in $u(F)$.

$$\pi_{a_0,\mathcal{M}}(r') = \min_{(p,a,r)\in\mathcal{M}} \sigma_{\mathcal{Q}\times\mathcal{A}}((p_0,a_0),(p,a)) \rightsquigarrow \sigma_{\mathcal{R}}(r,r') \tag{7.44}$$

for all $r' \in \mathcal{R}$, where $\rightsquigarrow$ is a generalized implication operator. An alternative approach using conjunction-based (example-based) fuzzy rules has been outlined in Chapter 5. It leads to the possibility distribution[22]

$$\pi_{a_0,\mathcal{M}}(r') = \max_{(p,a,r)\in\mathcal{M}} \min\{\sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)), \sigma_{\mathcal{R}}(r,r')\}. \tag{7.45}$$

Suppose that outcomes are directly given in terms of utilities, i.e. $U = \mathcal{R}$. (Otherwise, a possibility distribution on the set of utility degrees can be obtained via $v \mapsto \max_{r\,:\,u(r)=v} \pi_{a_0,\mathcal{M}}(r)$.) The problem of choosing an act then turns out as one of choosing among the possibility distributions

$$\{\pi_{a_0,\mathcal{M}} \,|\, a_0 \in \mathcal{A}\}. \tag{7.46}$$

This situation is quite similar to decision under risk where the agent has to choose among probability distributions (lotteries).

There are different ways of realizing a corresponding selection. We can, for instance, adopt a quantitative point of view and interpret possibility degrees as upper probabilities. A possibility distribution then corresponds to a special type of plausibility measure, which means that the methods discussed in Section 7.5.2 can be applied.

We can, however, also interpret the possibilistic approach in a purely qualitative way. Dubois and Prade [123] have recently proposed a qualitative decision theory in which uncertainty and utility are represented by possibility measures and qualitative utility functions, respectively. The corresponding decision criteria are derived from an axiomatic framework which can be seen as a qualitative counterpart to the axioms of Von Neumann and Morgenstern's expected utility theory.

Let $\sqsubseteq$ be a preference relation on the class $\Pi$ of normalized possibility measures on a finite set $\mathcal{R} = \{r_1, \ldots, r_n\}$ of outcomes. As usual, denote by $\sim$ and $\sqsubset$ the symmetric and anti-symmetric part of $\sqsubseteq$, respectively. Moreover, let $V$ be a finite linear scale of uncertainty such that $\min V = 0$ and $\max V = 1$. Likewise, let $U$ be a finite linear scale of preference such that $\min U = 0$ and $\max U = 1$. The commensurability between the ordinal scales $U$ and $V$ is achieved via an order-preserving mapping $h$ from the plausibility scale to the preference scale which satisfies $h(0) = 0$ and $h(1) = 1$. For $\lambda, \mu \in V$ with $\max\{\lambda, \mu\} = 1$ the possibilistic mixture $(\lambda/\pi, \mu/\pi')$ of two possibility distributions $\pi$ and $\pi'$ again defines a possibility distribution:

$$\forall\, r \in \mathcal{R} \,:\, (\lambda/\pi, \mu/\pi')(r) \overset{\mathrm{df}}{=} \max\{\min\{\lambda, \pi(r)\}, \, \min\{\mu, \pi'(r)\}\} \ .$$

In [103], the following axiomatic system P has been proposed:

[22] Note that in Chapter 5 this distribution has been denoted by $\delta$ instead of $\pi$. Here, this distinction is not needed.

P1 $\sqsubseteq$ is a total preorder.

P2 $\pi \leq \pi' \Rightarrow \pi' \sqsubseteq \pi$ (uncertainty aversion).

P3 $\pi_1 \sim \pi_2 \Rightarrow (\lambda/\pi_1, \mu/\pi) \sim (\lambda/\pi_2, \mu/\pi)$ (independence).

P4 $\forall \pi \in \Pi \; \exists \lambda \in V \; : \; \pi \sim (1/r^*, \lambda/r_*)$, where $r^*$ and $r_*$ denote a maximal and a minimal element of $\mathcal{R}$, respectively.[23]

Based on this set of axioms, the existence of a utility function $u : \mathcal{R} \longrightarrow U$ and the following *pessimistic* decision criterion, which represents the preference relation $\sqsubseteq$, are derived:

$$\mathrm{QU}^-(\pi) \overset{\mathrm{df}}{=} \min_{r \in \mathcal{R}} \max \{n(h(\pi(r))), u(r)\} \; . \tag{7.47}$$

That is $\pi \sqsubseteq \pi' \;\Leftrightarrow\; \mathrm{QU}^-(\pi) \leq \mathrm{QU}^-(\pi')$. Here, $n$ is the order-reversing function on $U$.

As an alternative model, an axiomatic system O has been proposed in which the uncertainty aversion axiom P2 is replaced by an uncertainty-prone postulate. Moreover, P4 is slightly modified:

O1 $\sqsubseteq$ is a total preorder.

O2 $\pi \leq \pi' \Rightarrow \pi \sqsubseteq \pi'$.

O3 $\pi_1 \sim \pi_2 \Rightarrow (\lambda/\pi_1, \mu/\pi) \sim (\lambda/\pi_2, \mu/\pi)$.

O4 $\forall \pi \in \Pi \; \exists \lambda \in V \; : \; \pi \sim (\lambda/r^*, 1/r_*)$.

Based on these axioms one obtains the *optimistic* decision criterion

$$\mathrm{QU}^+(\pi) \overset{\mathrm{df}}{=} \max_{r \in \mathcal{R}} \min \{h(\pi(r)), u(r)\} \; . \tag{7.48}$$

If we interpret the approach to CBI outlined in Chapters 5 and 6 as purely qualitative ones (and also assure the commensurability of the plausibility scale and the preference scale), the decision theory of [123] can be applied to the distributions (7.44) or (7.45). That is, the decision criteria derived from the above axioms can be used in order to choose the most preferred distribution from the set (7.46), and, hence, the most preferred act. Applying (7.47) resp. (7.48) leads to the following valuations of an act $a \in \mathcal{A}$:

$$\mathrm{V}^{\downarrow}(a) = \mathrm{QU}^-(\pi_{a,\mathcal{M}}) = \min_{r \in \mathcal{R}} \max \{n(h(\pi_{a,\mathcal{M}}(r))), u(r)\}, \tag{7.49}$$

$$\mathrm{V}^{\uparrow}(a) = \mathrm{QU}^+(\pi_{a,\mathcal{M}}) = \max_{r \in \mathcal{R}} \min \{h(\pi_{a,\mathcal{M}}(r)), u(r)\} \; . \tag{7.50}$$

Let us finally mention that the uncertainty averse and uncertainty prone postulates P2 and O2 can be replaced by (intuitively plausible and somewhat more

[23] We extend $\sqsubseteq$ to $\mathcal{R}$ in the usual way: $r \sqsubseteq r'$ iff $\pi_r \sqsubseteq \pi_{r'}$, where $\pi_r = \mathbb{I}_{\{r\}}$ and $\pi_{r'} = \mathbb{I}_{\{r'\}}$.

appealing) *possibilistic dominance* criteria which are possibilistic counterparts to the well-known concept of probabilistic dominance. This result is proved in Appendix A.

## 7.6 CBDM models: A discussion of selected issues

In this section, we shall discuss some selected issues in case-based decision making. Our emphasis is on pointing out some principal differences between the models outlined in previous sections. In order to demarcate the different approaches, we shall reserve the acronym CBDM mainly for the framework presented in Section 7.5. Since the methods from Sections 7.1–7.4 are closer to case-based decision theory originally introduced by Gilboa and Schmeidler, they will be referred to as CBDT.

### 7.6.1 The relation between similarity, preference, and belief

A main difference between the models outlined in Section 7.5 (CBDM) and the approaches of previous sections concerns the way in which the concepts of belief, preference, and similarity are related. The approaches of Sections 7.1–7.4 make use of a decision-theoretic setup which is based on the concepts of similarity and utility alone. As opposed to this, the framework of CBDM in Section 7.5 makes also explicit the concept of belief and can thus be seen as an extension of classical (statistical) decision-theoretic models. In fact, this approach realizes a two-stage process, in which the actual decision problem is only solved in the second stage by means of (more or less) common techniques from decision theory. Case-based reasoning is not used for selecting an act directly. Rather, it has influence on the formation of the *belief* of the decision maker. This belief is represented in the form of a belief function or possibility distribution on the set of outcomes, $\mathcal{R}$. The cases contained in a memory $\mathcal{M}$ are treated as observations. For instance, observing that an act $a$ has led to a good result for a similar problem will increase the agent's belief that $a$ is also a good choice for the problem at hand.

The derivation of (7.2) in [167] shows that an agent with a utility function $u$, who obeys the respective axioms, behaves *as if* it had a similarity measure over $\mathcal{Q}$ and evaluates acts according to (7.2). This way, similarity is directly related to utility and indirectly to preference. The formal resemblance of (7.2) and the EUT formula, i.e., the expected utility of an act, suggests that the meaning of similarity in CBDT is to some extent comparable to the role that probability plays in EUT.

Most approaches to decision making evaluate acts by combining preference and belief in some way, where preference is quantified in the form of a utility function. In fact, for estimating the utility one obtains when choosing a certain act it seems natural to consider the set $V$ of possible utility degrees,[24] to modify each degree $v$ in accordance with an associated degree of belief, and to aggregate these modified utilities.[25] In expected utility theory, for instance, degrees of belief associated with $v \in V$ (and an act $a$) correspond to probabilities $p_a(v)$, and modification and aggregation are realized by multiplication and addition, respectively:

$$V(a) = \sum_{v \in V} p_a(v) \cdot v. \tag{7.51}$$

Within the qualitative approach proposed in [103, 123], belief is represented by possibility degrees $\pi_a(v)$, modification corresponds to bounding the impact of less possible utility degrees upon the valuation of an act, and the min-operator is used as an aggregation function:

$$V(a) = \min_{v \in V} \max\{1 - \pi_a(v), v\}. \tag{7.52}$$

Observe that the averaged similarity version of (7.2) corresponds to the expected utility model (7.51) if the probability $p_a(v)$ is estimated according to

$$p_a(v) = \frac{\sum_{(p,a,r) \in \mathcal{M}, u(r)=v} \sigma_{\mathcal{Q}}(p, p_0)}{\sum_{(p,a,r) \in \mathcal{M}} \sigma_{\mathcal{Q}}(p, p_0)}. \tag{7.53}$$

Likewise, (7.19) is equivalent to (7.52) with

$$\pi_a(v) = \max_{(p,a,r) \in \mathcal{M}, u(r)=v} \sigma_{\mathcal{Q}}(p, p_0). \tag{7.54}$$

As can be seen, based on the idea that similarity is used for assessing a degree of belief, namely (7.53) resp. (7.54), it is possible to interpret the approaches (7.2) and (7.18) within an extended decision-theoretic framework which combines similarity, preference, and belief, even though the latter only appears implicitly.

Still, there are several motivations for modeling the (causal) relation between similarity and belief in a more explicit way, as we have done in Section 7.5. Firstly, viewing the cases of a memory as an (additional) information source which has an effect on the agent's belief and, hence, utilizing case-based reasoning for decision making only indirectly leads to a more expressive approach which also avoids some technical difficulties. This becomes obvious, for instance, when considering the extreme example of a memory that does not contain any case similar to the current problem, which means that the memory is effectively empty. If, however, no cases exist, it seems somewhat peculiar that a *case-based* (similarity-based) reasoning procedure should be used for estimating the utility of choosing some act for solving the problem. Instead of assigning a "default utility," it appears more natural

[24] For the sake of simplicity suppose this set to be finite.

[25] Note that the consideration of single utility degrees may not be enough if belief is formalized by means of non-additive measures of uncertainty [166, 330, 334].

to expect the result of case-based (similarity-based) reasoning to be *complete ignorance* about utilities, which is adequately reflected, e.g., by the possibility distribution $\pi \equiv 1$ on the set of outcomes. Needless to say, an uncertainty measure like a probability distribution, a belief function or a possibility distribution, is able to reproduce certain characteristics of a memory $\mathcal{M}$ better than a "point estimation." The averaged similarity version of (7.2), for instance, can be seen as a kind of weighted mean. It is unable, however, to represent the *variance* of utility degrees associated with a certain act.

Secondly, making uncertainty related to decision problems explicit allows for taking the agent's attitude toward uncertainty into account. Otherwise, this attitude has to be encoded in the similarity measure or the utility function. Suppose, for example, that a decision maker (repeatedly) faced with a problem $p$ can choose between two acts $a$ and $b$. Act $a$ yields utility 0 with certainty. The more risky act $b$ yields either an extremely high utility $M$ or an extremely low utility $-M$, where the high utility occurs with a fixed but unknown probability every time $b$ is chosen. The willingness of an (uncertainty averse) agent to choose $b$ will then depend on the number of times the cases $(p, b, M)$ and $(p, b, -M)$ have been observed.[26] The memory $\mathcal{M} = \{(p, b, M)\}$ containing only one case, for instance, might not be convincing enough, even though $V(b) = M > 0 = V(a)$ according to (7.2).

Thirdly, the distinction between two "mental" levels, one for representing knowledge and one for making decisions, seems to have advantages with respect to the design of intelligent systems [130], i.e., if a decision-theoretic model is understood as a language for modeling the problem solving behavior of an agent. The integration of different information sources at the decision level, for instance, would require a related extension of the underlying decision theory and seems to be more difficult than doing the same at the knowledge representation level. Consider again the approach of Section 7.5.3 as an example. There, case-based reasoning takes place at the knowledge representation level and yields a possibility distribution on the set of outcomes. It is hence not difficult to combine this case-based knowledge with general background knowledge represented, e.g., in the form of fuzzy rules. In fact, the possibility distributions associated with such rules can simply be combined (via intersection) with the distribution(s) originating from CBI.

Let us finally mention that the (causal) relation SIMILARITY → BELIEF is also supported by psychological evidence. In fact, the finding that people rely on similarity as a heuristic principle for assessing the *probability* of an uncertain event or the value of an uncertain quantity was made by TVERSKY and KAHNEMAN in various psychological studies [374]. The authors call this heuristic approach the *representativeness principle*. For example, the probability that a person has a certain job seems to be assessed by the degree to which this person is similar to the stereotype of a person having this job.

[26] In other words, the agent estimates the unknown probability that $M$ occurs by the corresponding (relative) frequency of occurrence.

### 7.6.2 The effect of observed cases

The impact that case-based information has upon the evaluation of acts is rather different for the decision models discussed in this chapter. A major difference concerns the question whether experienced cases can be compensated by other cases, e.g., whether a good experience can compensate for a bad one, or whether several moderately similar cases can outweigh one completely similar case.

Let us consider the last point first. Due to the accumulation of utility degrees in (7.2), a good experience with a very similar problem can be compensated by several good experiences with less similar problems. This contrasts, e.g., with the Nearest Neighbor decision rule (7.8) which takes only one (namely the most similar) observed case into account, i.e., which fully relies on the most relevant experience. Needless to say, the adequacy of the two principles will strongly depend on the application or, more precisely, the extent to which experience with a certain act can be transferred from one problem to a similar one. Consider, for instance, a medical agent having to choose between treatments $T_1$ and $T_2$. The successful application of therapy $T_1$ to several diseases with somewhat similar symptoms will generally not compensate for $T_2$'s curing exactly the same symptoms, even if $T_2$ has not been applied to any other disease.

Now, consider the second point, i.e., the question whether good experiences can compensate for bad ones and vice versa. The CBDT decision rule (7.2) as well as the averaged similarity version (7.28) do obviously allow for such a compensation, and the same is true for the Nearest Neighbor decisions in Section 7.2. As opposed to this, the criteria (7.19) and (7.20) compensate in only one direction: According to (7.19), an observed case can only *decrease* the evaluation of an act, which reflects the pessimistic or cautious character of this decision rule. Consequently, a positive experience cannot compensate for a negative one. Contrariwise, each observation can only positively influence the evaluation according to (7.20), i.e., a good experience is never annulled by a bad one.

Again, different evaluation principles will be adequate for different applications. In this connection, it should be noted that (7.19) and (7.20) assume an ordinal setting, whereas the addition and multiplication operators used by the CBDT criteria (7.2) and (7.3) make sense only for cardinal utility and similarity functions. Indeed, (7.19) and (7.20) might be preferred whenever it is difficult to define such functions. Consider again an example from the medical domain: Treatments $T_1$ and $T_2$ usually have the same effect. On the one hand, $T_2$ is less expensive than $T_1$. On the other hand, it is also more risky in the sense that is has already caused the death of some patients, whereas $T_1$ cures the disease with certainty. In this situation, it will of course be difficult to come up with a reasonable utility function, or to fix a minimal success rate of $T_2$ as a decision criterion.[27] Rather, one will generally be cautious and decide in favor of $T_1$, a decision behavior which is perfectly in line with (7.19).

[27] Extreme examples of this kind are often raised against expected utility theory.

It was pointed out above that (7.19) and (7.20) reflect very opposite attitudes of a decision maker. Let us finally mention that a similar remark applies to the indirect evaluations which deduce the *possibility* of outcomes. When using the criterion (7.49) in connection with the possibility distribution (7.44), the decision maker considers all outcomes as being fully possible as long as it has not made any observations. Each new case serves as a constraint and decreases the possibility of certain results. By applying the same decision rule to (7.45), the decision maker starts with the possibility distribution $\pi \equiv 0$.[28] Each new case serves as evidence for certain results and increases the possibility correspondingly. Loosely speaking, the agent learns what *can* happen, whereas it learns what *can* or *should not* happen if it relies on (7.44). The difference between (7.44) and (7.45) becomes also clear if we realize that (7.44) is based on the idea of an implication-based fuzzy rule, whereas (7.45) is related to the concept of a possibility rule, i.e., an example-based (conjunction-based) fuzzy rule.

### 7.6.3 Dynamic aspects of decision making

Since CBI is closely related to the idea of repeated problem solving and aspects of learning it seems natural to consider a CBDM agent acting over time in a certain environment. The question, then, is how successful a CBDM strategy proves to be. Since the acquisition of experience in the form of cases is an inherent part of CBDM, investigating a CBDM strategy in the context of repeated problem solving seems to be the only reasonable way of judging its efficiency.[29] Such an analysis, which will have much in common with the analysis of heuristic problem solving methods [291], is principally possible. For instance, given (among other things) the precise specification of a stochastic environment in which the agent acts as well as the specification of utilities of histories (which correspond to paths in this environment), the *expected* performance of a CBDM strategy is well-defined (cf. Section 7.4).

Let us mention, however, an interesting aspect of CBDM which makes the analysis of a given strategy as well as the selection of an optimal strategy difficult. Namely, a single decision at a certain point of time does not only affect the expected utility and future states of an agent directly. Rather, it has also an impact on the agent's *experience* and, hence, changes its future decision behavior. For the analysis of a given decision strategy this means that it has to take the (expected) evolution of the agent's memory into account. For the development of an optimal strategy it means that a single decision should not only be judged on the basis of some estimated (immediate) utility. Since a more experienced agent

[28] Again, note that this is actually a distribution of *guaranteed possibility*, denoted by the symbol $\delta$ in Section 5.

[29] Considering something such as the performance of a decision strategy makes sense if we concede to CBDM a *normative* character in connection with the idea of heuristic problem solving. Other criteria become relevant if a case-based decision theory serves as a *descriptive* theory of (human) decision making [171].

will probably make better decisions in the future it should also take the aspect of broadening experience into account. Informally speaking, the agent has to find a tradeoff between the *exploitation* of its past experience and the *exploration* of new decisions. This, of course, requires some kind of metalevel reasoning quite comparable to the concept of metalevel rationality in connection with expected utility theory (cf. the remarks on page 253). The aforementioned exploration–exploitation tradeoff is also well-known in fields like optimization and machine learning.

The above idea can be illustrated nicely for the model (7.2). As an interesting consequence of this decision principle it has been pointed out in [167] that it can be seen as a theory of "bounded rationality" formalizing SIMON's idea of "satisficing" [260, 344]. Suppose, for example, that the selection of a certain act $a_*$ for a problem $p$ has led to a positive utility $u(r(p, a_*))$. When faced with the same problem again, the decision maker will prefer this act to all other acts to which the default utility 0 is assigned (since they have not been tried yet). More generally, the agent may try several acts until one results in a positive utility, but it will not attempt to maximize utility. Now, an intuitively reasonable modification of the decision behavior prescribed by (7.2) is to try a new act $a$ from time to time. This way, an act $a^*$ such that $u(r(p, a^*)) > u(r(p, a_*))$ might eventually be found. Since the agent will then go on choosing $a^*$, this would have a positive impact on its (future) "welfare," a prospect that justifies to put up with the risk of sometimes obtaining a smaller utility. See [168] for a related strategy of realizing an "experimenting" agent in CBDT.[30] The idea is to adapt the aspiration level $\alpha$ in the generalization

$$V_{p_0,\mathcal{M}}(a) \overset{\text{df}}{=} \sum_{(p,a,r)\in\mathcal{M}} \sigma_{\mathcal{Q}}(p, p_0) \cdot (u(r) - \alpha)$$

of (7.2) by choosing an act at random from time to time. This way, the agent can avoid to get stuck in a suboptimal strategy.[31]

## 7.7 Experience-based decision making

Case-based decision making, as presented in different versions in previous sections, can basically be seen as a two-step procedure:

I. Estimation/evaluation: Given a set of experiences in the form of triples $(p, a, u)$ and a new problem $p_0$, one estimates the utility $u(p_0, a_0)$ for each act $a_0 \in \mathcal{A}$.

[30] As the "conservative" decision strategy of always choosing the act "go to a restaurant which has not been tried yet" shows, a careful distinction between the agent's decisions and its actual behavior has to be made. Particularly, a satisficing decision strategy does not necessarily entail conservative behavior.

[31] The same idea is also present in several approaches to learning in game theory.

II.Decision: An (apparently) optimal act is then chosen on the basis of these estimations.

It deserves mentioning that one actually has to distinguish between the *estimation* of a utility degree and the *evaluation* of an act. As opposed to (7.3), for instance, the value $V(a_0)$ in (7.2) is obviously not an estimation of the utility $u(p_0, a_0)$ but still an evaluation of the act $a_0$. As can be seen, many possibilities of act evaluation exist, although the estimation of an induced utility might be regarded as the most natural one. Subsequently, we make the reasonable assumption that the agent bases its decision on estimations of the utilities of acts $a_0$. That is, we assume that the agent is an *estimated* utility maximizer, just like a decision maker applying EUT is an *expected* utility maximizer.

Experience-based decision making [196] generalizes case-based decision making in the sense that the estimation of utility degrees as part of the above two-step procedure is realized by any learning method, not necessarily a case-based one. Note that EBDM thus defined is an *indirect* approach in which an approximation $\varDelta$ to the *optimal decision function*

$$\varDelta^* : \mathcal{P} \longrightarrow \mathcal{A}, \tag{7.55}$$

which maps problems to (optimal) decisions, is derived from an estimation $\hat{u}$ of the utility function $u : \mathcal{P} \times \mathcal{A} \longrightarrow \mathfrak{R}$:

$$\varDelta : d \mapsto \arg\max_{a \in \mathcal{A}} \hat{u}(p, a).$$

An obvious alternative is to realize EBDM in a more *direct* way. In this case, the agent tries to learn the decision function (7.55) directly, without estimating the utility function as an intermediate step.[32] This kind of EBDM, which appears especially appealing from an efficiency point of view, will be discussed in detail in Section 7.7.1.

Case-based decision making, as case-based reasoning in general, is closely related to learning from experience in the form of examples or facts. Investigating the link between factual knowledge and beliefs derived from that knowledge, this relation is also emphasized in [173], where the axiomatic foundation of CBDT is developed in a more general context, not restricted to decision making.

In the field of machine learning, several standard types of learning problems are distinguished. In this connection, it is interesting to note that the indirect approach to EBDM (as realized by CBDT) is closely related to *reinforcement learning*, at least from a formal point of view. There are different settings for reinforcement learning, most of which fall back on concepts from Markov Decision Processes: The decision making agent acts in some unknown environment defined by a set of states $S$. At each point of time, the agent finds itself in a state $s \in S$, where an action has to be performed. The consequences of performing action $a$ in

[32] Such agents are often called *reflex agents* in artificial intelligence.

state $s$ are determined by a *reward function*, $r$, and a *transition function*, $\delta$: The agent receives an immediate reward,[33] $r(s,a)$, and moves from state $s$ to state $\delta(s,a) \in S$. This process is repeated until eventually a *final state* is reached.

As can be seen, the notions of state and reward in reinforcement learning play the roles, respectively, of the concepts of problem and utility in CBDT. Moreover, the optimal decision function $\varDelta^*$ in EBDM basically corresponds to what is called an *optimal policy* in reinforcement learning. The basic difference between CBDT and reinforcement learning (sequential decision making) concerns the generation of problems resp. states. In sequential decision making, the next state is a (perhaps non-deterministic) function of the current state $s$ and the action $a$. Consequently, an action does not only determine the immediate reward, but also the next decision problem and, thereby, the prospect of future rewards. A "myopic" decision maximizing only the immediate reward $r(s,a)$ is hence not necessarily optimal. Rather, an optimal action should be one that maximizes the sum of the immediate reward and the (expected) future rewards.[34] A function taking this into account is the so-called $Q$-function, that can be defined as follows:

$$Q(s,a) = r(s,a) + \gamma \sum_{s' \in S} p(s,a,s') \cdot \max_{a'} Q(s',a'), \tag{7.56}$$

where $0 \leq \gamma \leq 1$ is a discounting factor. Moreover, $p(s,a,s')$ is the probability that act $a$ yields $s'$ as the successor state of $s$ (here the transition function $\delta$ is non-deterministic). This leads to Bellmann's optimality equations

$$V^*(s) = \mathbb{E}\left[ r(s, \varDelta^*(s)) + \gamma V^*(\delta(s, \varDelta^*(s))) \right], \tag{7.57}$$

which determine the optimal decision function $\varDelta^*$. Thus, the value of being in state $s$, $V^*(s)$, is the expected sum of the immediate reward and the discounted future rewards under optimal behavior, as prescribed by $\varDelta^*$.

In CBDT, the action chosen for a problem $p$ does not affect the occurrence of future problems, which are not under the control of the decision maker. Thus, an optimal decision is simply one that maximizes $u(p,a)$. Note that the same strategy, namely maximizing $r(s,a)$, is also optimal in sequential decision problems if either future rewards (utilities) are discounted by means of a discounting factor of $\gamma = 0$ or if the transition function $\delta$ (as realized, e.g., by the probability function $p$ in (7.56)) does not depend on $a$. In other words, there are two possibilities of viewing CBDT, at least formally, as a special type of reinforcement learning: Either the case-based decision maker follows a myopic strategy, or future states (problems) do not depend on actions.

The main objective in reinforcement learning is to estimate the Q-function on the basis of rewards obtained so far.[35] If $\widehat{Q}$ is such an estimation, an apparently

[33] In a more general version, feedback can also be delayed (e.g., the win or loss of a game).

[34] The addition of rewards might be replaced by an alternative aggregation operator, of course.

[35] The reward function $r$ and probability function $p$ in (7.56) are assumed to be unknown. Otherwise, classical approaches (e.g., dynamic programming techniques in the case of finite horizon decision problems) can be used to find an optimal policy on the basis of (7.57).

optimal strategy is given by the decision function

$$\Delta : s \mapsto \max_a \widehat{Q}(s,a).$$

Even though $\Delta(s)$ does indeed maximize the estimated rewards, the strategy of persistently choosing actions $\Delta(s)$ so as to maximize the current estimation of $Q$ is again somehow shortsighted and not necessarily optimal. Namely, the agent has to bear in mind that it keeps learning over time: The estimation of the $Q$-function is permanently revised in the light of new observations, and the improvement of $\widehat{Q}$ might be larger for an alternative action $a' \neq \Delta(s)$. Consequently, choosing $a'$ might lead to better decisions in the future, even though the immediate reward might be better for $a$. In other words, the agent has to find an *exploration–exploitation tradeoff* (cf. page 287): It has to trade off (estimated) rewards against the potential for learning useful new information. One strategy, for instance, is to make random choices, where the probability of an action is proportional to its estimated value. This way, preference is given to higher valued acts, but apparently suboptimal acts are not completely ignored. Another possibility is to assign relatively large default values to yet unknown states (or state–action pairs) so as to offer an incentive for exploring such states.

As already mentioned in Section 7.6.3, the exploration–exploitation problem is solved in a very similar way in CBDT, namely by assigning default utilities to problem–act tuples for which relevant experience is not available as yet. The induced satisficing behavior of a decision maker can be seen as a special exploration strategy: The agent tries new actions until a satisfying one has been found.

### 7.7.1 Compiled decision models

The modification (7.3) of evaluation (7.2) can be seen as a special version of estimated utility maximization as discussed in Section 7.7. In fact, (7.3) is nothing else than the application of Shephard's interpolation method [340], a special type of Nearest Neighbor (NN) estimation [76].[36] This method is well-known in machine learning, and it is used for making predictions in other CBR approaches as well (e.g., in the ELEM2-CBR system [61]).[37]

Of course, Shephard's interpolation method is not the only way of realizing the estimation step in EBDM. Principally, it could be replaced by any machine learning method. In this connection, it is especially interesting to distinguish between *instance-based* and *model-based* approaches to (supervised) machine learning [79]. In particular, our discussion in Section 2.1 has shown that instance-based learning, as a lazy approach, is easy and quite appealing from a knowledge revision and adaptation point of view, but not very efficient in the prediction phase. In

[36] This is the weighted $k$-Nearest Neighbor approximation with $k = n$, i.e., it makes use of the complete set of observations.

[37] See the paper [173] of Gilboa and Schmeidler for a comparison between the nearest neighbor method and their case-based approach.

the context of EBDM, this means that a case-based approach appears reasonable if the decision maker disposes of a *limited* number of experiences. If a large set of observed cases is available, however, a *compressed* representation of the agent's knowledge in the form of a *model* might be more efficient. The aspect of efficiency becomes especially relevant if deliberation time is costly or strictly limited as in real-time decision making [64]. Then, model-based learning might be preferred to instance-based learning, since it is the decision process itself rather than the learning process that is time critical.

In this connection, it is convenient to classify decision problems with respect to their novelty. According to a crude distinction, we can differentiate between problems which are solved frequently and hence become almost automated, problems for which deliberation is required but which are still familiar, and problems which are completely unfamiliar [167]. These problem types might be tackled most efficiently by means of different approaches to learning and knowledge representation:

- instance-based learning of the utility function for unfamiliar problems,
- model-based learning of the utility function for familiar problems,
- and a "compiled decision model" approach for routine decisions.

As already mentioned before, the idea of a compiled decision model is to learn the optimal decision function (7.55) directly, rather than making a detour by learning the utility function. In [325], *compilation* is understood as a method for omitting intermediate computations in some input–output relation. Thus, a compiled model is an execution architecture computing the original mapping, but doing so in a more efficient way. This approach will now be discussed in more detail.

When the decision maker tries to learn the utility function $u : \mathcal{P} \times \mathcal{A} \longrightarrow \mathfrak{R}$, a case $(p, a, u)$ can be considered as an example of the form $(x, u)$, where the input $x = (p, a)$ is a problem–act pair and the outcome $y = u(p, a)$ is a utility degree. Thus, learning the utility function fits the framework of *supervised* machine learning. Still, a case $(p, a, u)$ can also be interpreted in a different way, namely as a *valued example*. That is, $(x, y) = (p, a)$ is an example and $u$ an evaluation thereof. The target function is now the (optimal) decision function $\Delta^* : \mathcal{P} \longrightarrow \mathcal{A}$. Roughly speaking, the utility $u = u(p, a)$ indicates the quality of an associated example $(p, a)$.

The compiled model approach thus necessitates an extension of standard (supervised) learning methods which takes the valuation $u$ of an example $(p, a)$ into account. To this end, we shall fall back on the idea of "satisficing" as discussed above in connection with the model of CBDT: A "satisficing" decision maker discriminates between only two types of decisions, namely acceptable and non-acceptable ones. As will be seen below, the problem of inducing a (satisficing) decision model thus comes close to the standard setting of supervised learning.

### 7.7.2 Satisficing decision trees

In this section, we are going to propose a concrete approach to learning compiled decision models, namely a modified version of decision tree induction. Thus, the idea is to implement the decision function $\Delta : \mathcal{P} \longrightarrow \mathcal{A}$ as a decision tree[38] resp. a set of *condition–action rules* that yields as a classification the action to be chosen, given a new problem (condition). This approach is adequate if the set $\mathcal{A}$ of available acts is relatively small. Moreover, it assumes that problems are represented by attribute–value pairs with discrete-valued attributes.[39] Before presenting our approach, we give a brief introduction to decision tree learning.

The basic principle underlying most decision tree learners, well-known examples of which include the ID3 algorithm [304] and its successors C4.5 and C5.0 [306] as well as the CART system [55], is that of partitioning the set of given examples, $S$, in a recursive manner. Each inner node $\eta$ of the decision tree defines a partition of a subset $S_\eta \subseteq S$ of examples assigned to that node. This is done by classifying elements $s \in S_\eta$ according to the value of a specific attribute $T$. The attribute is selected according to a measure of effectiveness in classifying the examples, thereby supporting the overall objective of constructing a small tree.

A widely applied "goodness of split" measure is the *information gain*, $G(S,T)$, which is defined as the expected reduction in "impurity" (with regard to the class distribution) which results from partitioning $S$ according to $T$:

$$G(S,T) = I(S) - \sum_t \frac{|S_t|}{|S|} I(S_t), \tag{7.58}$$

where $S_t$ denotes the set of elements $s \in S$ whose value for attribute $T$ is $t$. Moreover, $I(\cdot)$ is a measure of impurity, such as the GINI function [55]

$$I(S) = \sum_{c \neq c' \in \mathcal{C}} q_c q_{c'} = 1 - \sum_{c \in \mathcal{C}} (q_c)^2 \tag{7.59}$$

with $q_c$ the proportion of elements $s \in S$ having class $c$. Besides, a number of alternative (im)purity measures, such as entropy, have been devised. See [268] for an empirical comparison of splitting measures.

Suppose a set $\mathcal{X}$ of instances to be given, where each instance is characterized by several attribute values. Moreover, each $x \in \mathcal{X}$ belongs to one class $c = \text{class}(x) \in \mathcal{C}$. Given a set of training samples $S = \{(x_1, c_1), \ldots, (x_n, c_n)\} \subseteq \mathcal{X} \times \mathcal{C}$, the basic ID3 algorithm derives a decision tree as follows:

- The complete set of training samples, $S$, is assigned to the root of the decision tree.

[38] Decision tree learning is actually a classification method. Even though classifications can be considered as decisions, it is not specifically used for decision making in the proper sense. Therefore, one might prefer the alternative terms discimination or classification tree.

[39] Continuous-valued attributes can be discretized before or during the learning of a decision tree [93].

- A node $\eta$ becomes a leaf (answer node) of the tree if all associated samples $S_\eta$ belong to the same class or if all attributes have already been used along the path from the root of the tree to $\eta$.
- Otherwise, node $\eta$ becomes a decision node: It is split by partitioning the associated set $S_\eta$ of examples. This is done by selecting an attribute as described above and by classifying the samples $s \in S_\eta$ according to the value for that attribute. Each element of the resulting partition defines one successor node.

Once the decision tree has been constructed, each path can be considered as a rule. The precedent part of a rule is a conjunction of conditions of the form $T = t$, where $T$ is an attribute and $t$ a specific value thereof. The consequent part determines a value for the class variable. New examples are then classified on the basis of these rules, i.e., by looking at the class label of the leaf node whose attribute values match the description of the example. Notice that a unique class label is associated with each answer node if the data is not noisy and, hence, the original sample does not contain any *clashes* (cases with identical attribute vectors but different classes). Otherwise, the distribution of class labels at the leaf can be used for deriving a probabilistic estimate. Quite often, the induced tree undergoes further (post-)processing [267]. Here, the objective is to *prune* large trees in order to guarantee transparency. Moreover, pruning counteracts the problem of overfitting.

An incremental decision tree algorithm has been proposed in [377]. Given the same training data, this algorithm induces the same tree as ID3. Now, however, instances are processed in a serial way, that is, the current decision tree is updated each time a new example arrives. Since algorithmic aspects are not our main concern, we refrain from describing the algorithm here. It should be noted, however, that an incremental approach to learning has considerable advantages, especially in the context of decision making. In fact, the idea of learning and improving performance over time is one of the major aspects of case-based or experience-based decision making.

In this connection let us also mention a method that combines decision tree learning and lazy learning [155]: Given a set of training data, new instances are classified by means of a decision tree. However, a new tree is built for each individual instance (as in lazy learning, the complete data thus needs to be stored). Loosely speaking, this algorithm induces decision trees which are optimal for the individual instances, whereas a usual decision tree is good *on average*. The algorithm is efficient due to the fact that actually only one path of a tree is constructed, namely the one needed to classify the new instance. Besides, computational efficiency is improved by means of a caching scheme.

Let $\mathcal{P} = T_1 \times T_2 \times \ldots \times T_m$ be a set of potential problems, where $T_\imath$ denotes the (finite) domain of the $\imath$-th attribute. Thus, each problem $p \in \mathcal{P}$ is represented as a vector $p = (t_1, \ldots, t_m)$ of attribute values. Moreover, let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_k\}$ be a set of available actions. Finally, utility degrees are again measured on the real number line.

Assume a memory of cases

$$\mathcal{M} = \big\{(p_1, a_1, u_1), \ldots, (p_n, a_n, u_n)\big\} \in \mathcal{P} \times \mathcal{A} \times \mathfrak{R}$$

to be given, where $u_\imath = u(p_\imath, a_\imath)$ is the utility that has resulted from applying act $a_\imath$ to problem $p_\imath$. Let $\mathcal{M}_\mathcal{P}$, $\mathcal{M}_\mathcal{A}$, and $\mathcal{M}_{\mathcal{P}\times\mathcal{A}}$ denote the projection of $\mathcal{M}$ to $\mathcal{P}$, $\mathcal{A}$, and $\mathcal{P} \times \mathcal{A}$, respectively. The idea pursued here is to compile this case base into a decision tree which is then used for solving future decision problems.

Let $u^* \in \mathfrak{R}$ be a utility threshold defined by the decision maker. This threshold corresponds to the aspiration level in the CBDT model of Gilboa and Schmeidler: An action $a$ is *acceptable* for a problem $p$ if $u(p, a) \geq u^*$, and it is not acceptable if $u(p, a) < u^*$.

In a first step, each case $(p_\imath, a_\imath, u_\imath)$ is changed into an *example* $(p_\imath, a_\imath)$. The latter is called a *positive example* if $u_\imath \geq u^*$ and a *negative example* if $u_\imath < u^*$. In a second step, a *modified memory*, $S^*$, is derived from $\mathcal{M}$. For each problem $p \in \mathcal{M}_\mathcal{P}$ it contains a *generalized example* $(p, A_p)$, where $A_p$ denotes the set of *feasible acts* for problem $p$. This set is defined as follows:

$$a \in A_p \Leftrightarrow \begin{cases} u(p,a) = u_{max}(p, \mathcal{M}) & \text{if } \; u_{max}(p, \mathcal{M}) \geq u^* \\ (p,a) \notin \mathcal{M}_{\mathcal{P}\times\mathcal{A}} & \text{if } \; u_{max}(p, \mathcal{M}) < u^* \end{cases},$$

for all $a \in \mathcal{A}$, where

$$u_{max}(p, \mathcal{M}) \overset{\text{df}}{=} \max_{(p,b) \in \mathcal{M}_{\mathcal{P}\times\mathcal{A}}} u(p, b).$$

In plain words, an action $a$ is feasible for $p$ if it belongs to the best among the actions known to be acceptable for $p$, or if no acceptable action is known and $a$ has not been tried as yet. Notice that $A_p = \emptyset$ if all available acts have been applied to $p$ but none of them was acceptable, that is, if an acceptable act for $p$ does actually not exist. In this case, the decision maker should reduce the utility threshold $u^*$.[40] Subsequently, we assume $A_p \neq \emptyset$ for all $p \in \mathcal{P}$.

Before proceeding, let us point to a meaningful weakening of the above feasibility condition $u(p, a) = u_{max}(p, \mathcal{M})$. In fact, this condition could be replaced by $u(p, a) \geq u_{max}(p, \mathcal{M}) - \varepsilon$, where $\varepsilon \geq 0$ is a tolerance threshold. Here, the idea is that an action is acceptable even if its utility is slightly below the utility of the best (known) action. Of course, a decision maker being less ambitious in this sense will usually be able to induce simpler decision functions, i.e., to gain efficiency at the cost of decision quality.

We are now ready to formulate a generalized version of the decision tree learning problem whose objective is to induce a decision tree that prescribes, for any

[40] It would also be possible to maintain individual thresholds for the problems.

problem, an acceptable action:[41] Given a set of generalized examples

$$S^* = \big\{(p_1, A_{p_1}), (p_2, A_{p_2}), \ldots, (p_n, A_{p_n})\big\} \subseteq \mathcal{P} \times 2^{\mathcal{A}}, \tag{7.60}$$

induce a decision tree which implements a decision function $\Delta : \mathcal{P} \longrightarrow \mathcal{A}$ such that

$$\forall\, p \in \mathcal{M}_{\mathcal{P}} \,:\, \Delta(p) \in A_p.$$

In this problem, splitting a set of examples (7.60) is no longer necessary if

$$A(S^*) = \bigcap_{\imath=1}^{n} A_{p_\imath} \neq \emptyset, \tag{7.61}$$

hence, (7.61) defines a natural stopping condition. The corresponding node $\eta$ in the decision tree then becomes a leaf, and any action $a \in A(S^*)$ can be chosen as the prescribed decision $a_\eta$ associated with that node.

The main modification concerns the "goodness of split" measure. Let $G(\cdot)$ denote the information gain (7.58) as used for classical decision tree learning. That is, $G(S, T)$ quantifies the quality of the split of a (standard) sample $S$ induced by the attribute $T$. Now let the class of *selections*, $\mathcal{F}(S^*)$, of the generalized set of examples (7.60) be given by the class of samples

$$\{(p_1, a_1), (p_2, a_2), \ldots, (p_n, a_n)\} \subseteq \mathcal{P} \times \mathcal{A}$$

such that $a_\imath \in A_{p_\imath}$ for all $1 \leq \imath \leq n$. We extend the measure $G(\cdot)$ to generalized samples $S^*$ as follows:

$$G(S^*, T) \overset{\mathrm{df}}{=} \max_{S \in \mathcal{F}(S^*)} G(S, T). \tag{7.62}$$

As can be seen, the extended measure (7.63) is the ordinary measure obtained for the most favorable instantiation of the generalized examples $(p, A_p)$ and hence defines a "potential" goodness of split. It corresponds to the "true" measure that would have been derived for the attribute $T$ if this instantiation is compatible with the ultimate decision tree. Taking this optimistic attitude is clearly justified since the tree is indeed constructed in an (apparently) optimal manner (hence averaging would hardly make sense).

Computing (7.62) comes down to solving a combinatorial optimization problem, namely to finding

$$I(S^*) \overset{\mathrm{df}}{=} \min_{S \in \mathcal{F}(S^*)} I(S) \tag{7.63}$$

for (sub-)sets $S^*$ of extended examples, where $I(\cdot)$ is a measure of impurity. It might hence be regarded as critical from a time complexity point of view,

[41] An alternative approach would be to learn, for any action, the class of decision problems to which it can be applied. This type of problem fits into the framework of multi-label classification in machine learning. However, it does not provide efficient condition-action rules.

especially in the context of real-time decision making. One should keep in mind, however, that not the construction (or revision) of the decision tree is time critical but rather its application. In fact, real-time decision making must not be confused with real-time learning; as in other decision support systems, learning will rather be realized as an "off-line" procedure [64].

In Appendix G, we present a heuristic search method for computing (7.63) based on a branch & bound technique. The efficiency of the method depends critically on the number of actions (which is the maximal branching factor of the search tree) and the average size of the sets $A_p$ (which determines the average branching factor). Even without a detailed complexity analysis, experience has shown that no problems occur if the number of actions is small. For example, for six actions and sample sizes up to $n = 500$ the generalized splitting measure can be computed within the bounds of seconds. Still, if the number of actions is too large, the exact computation of (7.63) becomes intractable.

As an alternative we therefore suggest the following heuristic approximation of (7.63):

$$I(S^*) \; = \; I(S_*), \tag{7.64}$$

where the selection $S_* \in \mathcal{F}(S^*)$ is defined as follows: Let $q_\imath$ be the frequency of the action $\alpha_\imath$ in the set of examples $S^*$, i.e., the number of examples $(p_\jmath, A_{p_\jmath})$ such that $\alpha_\imath \in A_{p_\jmath}$. The $\alpha_\imath$ are first "preferentially ordered" according to their frequency $q_\imath$ (ties are broken by coin flipping), starting with the most frequent one. Then, the most preferred action $\alpha_\imath \in A_{p_\imath}$ is chosen for each example $p_\imath$. Clearly, the idea underlying this selection is to make the distribution of labels as skewed (non-uniform) as possible, since distributions of this type are favored by the impurity measure. In [204], we found that the measure (7.64) yields very good results in practice and compares favorably with alternative extensions of splitting measures.

Let us finally mention that the adequacy of a decision tree as a representation of the decision function $\Delta$ does of course depend on the structure of the optimal decision function $\Delta^*$. In fact, since a decision tree, at least in its standard version, partitions the problem (attribute) space $\mathcal{P}$ by means of axis-parallel decision boundaries, good results (in terms of both complexity and accuracy) are to be expected only if $\Delta^*$ is at least approximately compatible with this type of inductive bias.

### 7.7.3 Experimental evaluation

In order to get an idea of how the satisficing decision tree approach performs we have employed a procedure that generates decision problems in a systematic way. A *decision environment* is specified by the following parameters:

– The number $m$ of attributes describing a decision problem.

- The number $k$ of possible actions.
- The number of values for each attribute. For the sake of simplicity, we assume that each attribute has the same number $v$ of values. Without loss of generality these values are represented by natural numbers, that is $T_\jmath = \{1, 2, \ldots, v\}$ for all $1 \leq \jmath \leq m$.
- The utility function $u$ that assigns a utility degree $u(p, a)$ to each problem $p \in \mathcal{P}$ and action $a \in \mathcal{A}$. The generation of $u$ is realized by a random procedure which is under the control of a complexity parameter $\gamma$, as described below.

It has already been mentioned that the adequacy of a decision tree representation depends on the structure of the decision function $\varDelta^*$. In fact, $\varDelta^*$ can be represented by small trees if its structure is in agreement with a decision tree-like partitioning of the feature space; otherwise, the decision tree model might become rather complex. In order to control the complexity of the decision environment we have generated a utility function $u$ as follows: In a first step, an optimal decision tree is generated at random. This is done in a recursive manner by starting with the root of the tree and deciding for each node whether it is an inner node or a leaf of the tree. The probability of a node to become an inner node is specified by a parameter $0 < \gamma < 1$. Each inner node at level $\imath$ has $v$ successors, each of which corresponds to a value of the $\imath$-th attribute. If the tree has been generated, each leaf $\eta$ covers a subset $\mathcal{P}_\eta$ of the set of problems $\mathcal{P}$, namely those problems which match the attribute values associated with $\eta$. The leaf node $\eta$ is then assigned an optimal decision $a_\eta$ at random. From the resulting optimal decision tree, the utility function is finally derived by letting $u(p, a) = 1$ if $a$ is the optimal solution to $p$, i.e., if $a = a_\eta$, where $\eta$ is the leaf node that covers $p$. For all other (suboptimal) actions $b$ the utility $u(p, b)$ is defined as a decreasing function of the distance between $a$ and $b$ (where the distance between act $a_\imath$ and act $a_\jmath$ is $|\imath - \jmath|$). Note that the complexity is completely determined by the parameter $\gamma$: The larger $\gamma$, the larger the expected size of the optimal decision tree, i.e., the more complex the decision environment (at least for an agent that employs a decision tree representation of its decision model).

After having specified a concrete decision environment by generating the utility function $u$ at random, a simulation experiment is performed as follows: At the beginning, the memory of the decision maker is empty, and its (satisficing) decision tree corresponds to a single node. In the $\imath$-th decision epoch, a decision problem $p_\imath$ is chosen at random from $\mathcal{P}$, according to a uniform distribution. For this problem, the decision maker selects an action $a_\imath$ according to the current decision tree model (ties are broken by coin flipping). The new case, consisting of the problem $p_\imath$, the act $a_\imath$, and the experienced utility $r_\imath = u(p, a)$, is added to the memory. Moreover, the satisficing decision tree is updated whenever necessary. The simulation stops after the $L$-th decision epoch.

**Illustrating example.** To illustrate, we present a simple example step by step. Let $m = 6$, $k = 3$, $v = 3$, $L = 10$, $u^* = 0.7$ and consider the sequence of decision problems in Table 7.1 (and disregard, for the time being, the actions and utility degrees shown in the same table).

| problem | | | | | | act | utility |
|---|---|---|---|---|---|---|---|
| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | | |
| 2 | 1 | 2 | 1 | 1 | 2 | $\alpha_1$ | 1 |
| 2 | 1 | 2 | 3 | 3 | 1 | $\alpha_1$ | 1 |
| 1 | 1 | 2 | 3 | 2 | 3 | $\alpha_1$ | 0.2 |
| 1 | 3 | 3 | 3 | 3 | 3 | $\alpha_2$ | 0.8 |
| 1 | 1 | 3 | 2 | 1 | 3 | $\alpha_2$ | 1 |
| 2 | 2 | 1 | 2 | 1 | 3 | $\alpha_1$ | 0.9 |
| 3 | 2 | 2 | 2 | 2 | 3 | $\alpha_1$ | 0 |
| 3 | 1 | 2 | 2 | 3 | 2 | $\alpha_2$ | 0.5 |
| 3 | 1 | 1 | 3 | 3 | 1 | $\alpha_3$ | 0 |
| 1 | 3 | 2 | 1 | 1 | 2 | $\alpha_1$ | 0.8 |

**Table 7.1.** Sequence of decision problems specified by the values of six attributes (columns 1–6), the action performed by the decision maker, and the resulting utility degree.

For the first decision problem, the agent chooses an action at random. As shown in Table 7.1, this is action $\alpha_1$, for which it receives a utility of 1. Thus, the agent generates a decision tree which corresponds to the following rules:

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | action |
|---|---|---|---|---|---|---|
| ? | ? | ? | ? | ? | ? | $\alpha_1$ |

This tree prescribes action $\alpha_1$ regardless of the attribute values. Thus, for the second problem the agent chooses again $\alpha_1$, and again obtains a utility of 1. Consequently, it does not modify the decision tree. For the third problem, however, $\alpha_1$ yields a utility of 0.2 which falls below the utility threshold $u^*$. Therefore, the agent changes the decision tree according to the procedure outlined above:

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | action |
|---|---|---|---|---|---|---|
| 1 | ? | ? | ? | ? | ? | $\alpha_2$ |
| 2 | ? | ? | ? | ? | ? | $\alpha_1$ |

This tree prescribes to choose $\alpha_1$ if the value of the first attribute is 2, but $\alpha_2$ if this value is 1. Thus, the agent's hypothesis is that $\alpha_1$ yields bad outcomes if $t_1 = 2$ (note that $t_5$ and $t_6$ might have been chosen as splitting attributes as well). The next update occurs after the 7-th problem. Since the decision tree does not prescribe an action for $t_1 = 3$, the agent chooses $\alpha_1$ at random. This leads to a utility of 0. The new decision tree contains the following rules:

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | action |
|---|---|---|---|---|---|---|
| 1 | ? | ? | ? | ? | ? | $\alpha_2$ |
| 2 | ? | ? | ? | ? | ? | $\alpha_1$ |
| 3 | ? | ? | ? | ? | ? | $\alpha_2$ |

This tree is changed into

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | action |
|---|---|---|---|---|---|---|
| 1 | ? | ? | ? | ? | ? | $\alpha_2$ |
| 2 | ? | ? | ? | ? | ? | $\alpha_1$ |
| 3 | ? | ? | ? | ? | ? | $\alpha_3$ |

after the 8-th problem, since $\alpha_2$ yields $u = 1/2 < u^*$ for a problem with $t_1 = 3$. Notice that, so far, the decision is completely determined by the first attribute. After the 9-th problem, however, the agent realizes that this is not enough. The new decision tree also involves attribute $t_6$:

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | action |
|---|---|---|---|---|---|---|
| 1 | ? | ? | ? | ? | 3 | $\alpha_2$ |
| 2 | ? | ? | ? | ? | 3 | $\alpha_1$ |
| 3 | ? | ? | ? | ? | 3 | $\alpha_2$ |
| ? | ? | ? | ? | ? | 1 | $\alpha_1$ |
| ? | ? | ? | ? | ? | 2 | $\alpha_1$ |

**General findings.** More generally, we were interested in effects of the complexity of the decision environment and of the aspiration level of the decision maker. Regarding the first factor, more complex decision environments are expected to entail larger decision trees and smaller average utilities over time. As concerns the aspiration level, it is to be expected that higher levels will probably guarantee higher utilities on average but, at the same time, lead to more complex decision models. To illustrate, consider the problem of choosing the optimal dose of a drug for different patients. The simple decision tree shown in Fig. 7.1 might lead to satisfying results (the utility of a decision depends on the patient's state of health after the treatment). Still, even better results might be obtained by differentiating more precisely between patients, taking further attributes such as weight into account. This would of course mean using a more complex decision tree.

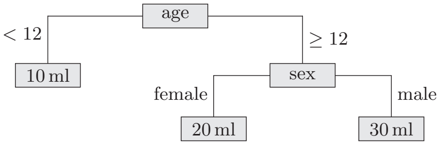


**Fig. 7.1.** Decision tree implementing a simple strategy for choosing the dose of a drug.

One might furthermore suspect that the effect of increasing the threshold $u^*$ is not independent of the complexity of the decision environment. Consider the following example: Suppose that exactly one optimal action with utility 1 exists for each problem. An act applied to a problem for which it is not optimal yields a utility of $0 < \alpha < 1$. Now, with a threshold $u^* \leq \alpha$ the decision maker is always satisfied, regardless of what action it applies to a problem. In fact, its decision model consists of only one rule which prescribes to choose act $a$, where $a$ is the act that has been applied to the first problem, perhaps randomly. The average utility is exactly $\alpha + (1-\alpha)/k$. If the utility threshold $u^*$ exceeds $\alpha$, the decision maker is satisfied only with the optimal acts, and it will spend enormous effort in finding these acts. The difficulty of this venture in turn depends on the generalization capability of the induced decision trees. If a decision tree is indeed a good model for the application at hand, the agent might succeed very quickly. Otherwise, it might try several actions for each individual problem before finally finding the optimal one it seeks for. Anyway, the decision model will become much more complex in this case. Of course, this model will finally come up with an optimal act for each problem. It should be noted, however, that it might take a long time and many unsuccessful attempts before this model is constructed. Therefore, the gain in utility might be poor over a limited time horizon and might hence not compensate for the increased complexity.

For different combinations of utility thresholds $u^*$ and complexity measures $\gamma$, we have performed 1,000 simulation runs with $m = 6$, $k = 4$, $v = 4$, $L = 100$, respectively. For each simulation, the average utility $(r_1 + \ldots + r_{100})/100$ and the average size of the decision tree have been computed. (The size was measured in terms of the number $l$ of leaf nodes.) The corresponding results, documented in Table 7.2 and Appendix H, permit the following conclusions which confirm our above suppositions: Increasing $u^*$ always leads to more complex decision models. However, it yields an improvement in average utility only if the decision environment is not too complex. Roughly speaking, the decision maker should be demanding for simple environments, where decision trees provide an adequate model and, hence, looking for better decision models is likely to be successful. If the environment is complex, however, it is urged to be modest: Searching for better models will generally increase the size of decision trees but hardly the quality of decisions.

| | $\gamma = 0.5$ | $\gamma = 1$ |
|---|---|---|
| $u^* = 0.2$ | 0.59 | 0.59 |
| $u^* = 0.9$ | 0.79 | 0.56 |

| | $\gamma = 0.5$ | $\gamma = 1$ |
|---|---|---|
| $u^* = 0.2$ | 8.84 | 15.76 |
| $u^* = 0.9$ | 19.82 | 48.79 |

**Fig. 7.2.** Average values of the (average) utility degrees (left) and (average) number of leaf nodes (right), taken over the 1,000 simulation runs.

## 7.8 Summary and remarks

### Summary

- We have briefly reviewed the original idea of case-based decision making due to Gilboa and Schmeidler (Section 7.1) as well as an alternative (fuzzy set-based) model proposed by Dubois and Prade (Section 7.3). Rather than concentrating on the accumulated or average performance of acts, the latter gives preference to acts which have always led to good results for problems which are similar to the current one.
- Methods of CBDM on the basis of the Nearest Neighbor principle have been investigated and characterized axiomatically in Section 7.2. NN decision rules can be seen as approximations of the decision criteria in CBDT. They can be motivated, among other things, for reasons of computational efficiency.
- The fuzzy set-based approach to CBDM has been generalized in Section 7.4. The extreme (worst case) valuation in the original model has been relaxed by looking out for acts which have yielded good results at least in *most* (rather than all) cases in the past. It has been shown that the relaxation of the "always" requirement in the principle underlying the original decision criterion can be advantageous in the context of repeated decision making.
- Section 7.5 has outlined an alternative CBDM framework. Corresponding methods combine results of previous chapters and generalized decision theories which have recently been proposed in literature in order to realize case-based decision making. These methods are *case-based* in the sense that an agent makes use of case-based reasoning (in the form of case-based inference) in order to support the modeling of a new decision problem, notably the specification of an uncertainty measure over possible outcomes. Since the latter is not necessarily a probability measure, the concept of an expected utility has to be generalized in order to compare acts. Two concrete methods have been discussed: The CBI approaches of Section 4.5 and Chapters 5 and 6 give rise to decision making with "case-based" belief functions and "case-based" possibility distributions, respectively.
- In Section 7.7, we introduced a framework of *experienced-based decision making* as an extension of case-based decision making. In EBDM, an agent faced with a new decision problem acts on the basis of experience gathered from previous problems in the past, either through predicting the utility of potential actions or through establishing a direct relationship between decision problems and appropriate actions. A realization of the latter approach has been proposed in the form of "satisficing decision trees".

## Remarks

- A representation of cases which is similar to the one proposed by Gilboa and Schmeidler was already suggested by Kolodner [234]. Apart from a problem and a solution she introduced a third component of a case: the *outcome* is thought of as the state of the world under the condition that the corresponding solution is applied and usually comprises some kind of feedback (see also [30]). According to this point of view, a triple $(p, a, r)$ is seen as an extended decription of a case, i.e., a usual case $(p, a)$ supplemented by some valuation $r$. By using the notation $\langle (p, a), r \rangle$ we have suggested a second interpretation in this chapter: $p$ and $a$ are taken together and constitute the first component of an *ordinary* case. This component is now partly under the control of the agent which can choose $a$. The second component is the outcome associated with the problem–act tuple $(p, a)$. Even though formally equivalent to the first notation, considering a case as a tuple $\langle (p, a), r \rangle$ seems more natural in the context of Section 7.5 where case-based reasoning (case-based inference) is used in its basic form, namely for predicting the outcomes associated with inputs (= problem–act tuples).
- In [172], Gilboa and Schmeidler provide an interesting comparison between case-based and rule-based knowledge representation, with special emphasis on the problem of induction. This article also contains further examples showing that the linearity of the CBDT functionals will often be too restrictive in practice. Particularly, this seems to be true if the decision maker is allowed to *learn* a similarity function resp. the importance of cases.[42] For instance, if experience is better represented by *subsets* of cases, the weight of an individual case depends on other observations as well. This effect, however, cannot be captured by the (additively) separable CBDT functionals but rather calls for the use of non-additive set-functions.
- As the summation of (weighted) degrees of *utility* in (7.2) reveals, CBDT actually assumes that the application of an act to similar problems yields similar utilities rather than similar outcomes. Of course, the two principles are only equivalent if outcomes are directly given in terms of utilities. Otherwise, the use of utility degrees in (7.2) has to be justified by the additional assumption that similar outcomes have similar utilities.
- The memory (7.1) of cases represents the experience of the decision maker. This does not mean, however, that all cases have been collected by the agent itself, or that the agent has made all related decisions. In fact, cases can also be experienced in a passive way or might even be the product of some kind of hypothetical reasoning.
- The simple accumulation of utility degrees in (7.2) does not always appear plausible, of course. Let us mention, however, that it might well be reasonable in connection with certain applications, such as the modeling of consumer behavior

[42] The adaptation of the similarity function is interpreted as some kind of *second-order induction* in [172].

in economics [170]. Interestingly enough, some undesirable effects of the accumulative nature of (7.2) can also be avoided by using the more general approach (7.5): The relevance of an observation might be reduced if the same case has already been encountered before. This idea seems quite plausible from a cognitive point of view. In fact, it again reveals the advantage of a "relevance-based" decision theory which is more general than a "similarity-based" approach.

- It has been mentioned that CBDT should not be seen as a competing theory, but as an alternative model which complements expected utility theory in a reasonable way. The claim that neither of them is superior in general and that the adequacy of a model strongly depends on the kind of problem under consideration is supported by a theoretical result of Matsui [262]. He shows that EUT and (a slight modification of) Gilboa and Schmeidler's CBDT are equivalent in the sense that each EUT model can be represented in the framework of CBDT and vice versa.[43] The embedded model, however, might be much more complex than the original model.
- Notwithstanding the cognitive appeal of CBDT, one might feel some uneasiness concerning the manifold possibilities for defining a case-based decision model. CBDT basically suggests that the current decision is a function of the agent's *experience*, considered against the background of a similarity relation between inputs. The experience, as represented by the history of cases, is an element of a quite complex and high-dimensional space on which various decision functions can be defined. Moreover, similarity is a rather vague concept, and it is by no means obvious what a reasonable similarity function should look like. In this respect, expected utility theory appears more restrictive. In fact, a decision is derived from a utility function[44] and a probability function which can be seen as an (information-compressed) *statistic* of the agent's experience (at least if probabilities are obtained from relative frequencies). Besides, the linear combination of probability and utility by means of the expected utility formula seems more straightforward than a similarity-based evaluation. Loosely speaking, EUT determines the information to be extracted from the agent's experience and the way in which this information is to be used more strictly.
- We have assumed the nearest neighbor (7.11) to be unique. The case of non-uniqueness could be handled by means of a set-valued generalization in the Dempster-Shafer style. Then, (7.11) defines the set of nearest neighbors, thus playing a role somewhat comparable to a focal element of a belief structure over $\mathcal{A}$. (Observe, however, that the weights in (7.10) are utility degrees which do not necessarily sum up to 1.) Moreover, (7.10) becomes

[43] Consequently, the two theories are *observationally* equivalent.

[44] Note that a utility function is principally required in CBDT as well. There, however, the function needs to be known only partially, namely for the observed outcomes.

$$V(a_0) = \sum_{(p,a)\in\mathcal{M}^{\downarrow}:\mathsf{NN}_{p_0,\mathcal{A}}(p,a)\ni a_0} \sigma_{\mathcal{Q}\times\mathcal{A}}((p,a),(p_0,a_0)) \cdot u(r(p,a)),$$

i.e., $V(a_0)$ defines the counterpart to the plausibility of a value $a_0 \in \mathcal{A}$.

- In the context of CBDM, the decision maker treats an uncertainty measure derived via CBI as some kind of "intermediate result" of the complete decision procedure. In Section 7.5.3, for instance, the possibility distributions (7.46) are taken as primitives in the second step of this procedure, namely the ranking of acts according to (7.49) or (7.50). In order to apply these qualitative decision criteria, the agent has to consider the distributions as being objectively given. In fact, the axiomatic framework in [123] is set up in the style of Von Neumann and Morgenstern: A utility function is derived from preferences, but the concept of belief in the form of a possibility distribution on outcomes is assumed to be given.[45] The two-stage procedure realized by CBDM might appear vulnerable from this point of view, particularly since the meaning of objectivity seems less obvious in the case of a possibility distribution than in the case of a probability [194].
- The idea of relating similarity and uncertainty (cf. Section 7.6.1) is also realized in the theory of counterfactuals proposed by Lewis [250], where the plausibility of an imaginary input is determined by its similarity to the current input.
- A combination of the concepts of similarity, preference (utility), and belief (probability) has also been outlined in [319]. However, this approach is quite different from the ideas discussed in this chapter. Particularly, it is not related to case-based reasoning.
- In connection with NN decision rules (Section 7.2) it has been mentioned that a decision maker will generally not utilize its complete memory when having to perform a prompt action. This consideration reveals the importance of efficient memory organization and case retrieval strategies. Needless to say, a computationally efficient (and cognitively plausible) case-based decision theory has to take these aspects into account.
- The methods proposed in Section 7.5 are based on *generalizations* of expected utility theory. Let us mention that one could also think of other ways of combining case-based inference and EUT. The constraint-based approach to CBI discussed in Chapter 3, for instance, can be used in order to suggest a subset of acts or states of nature which should be taken into account. EUT can then be applied to the reduced setup. Not only is an approach of this kind computationally efficient, it also appears cognitively plausible. In fact, human decision makers will generally concentrate on a small number of acts and disregard states of nature which are considered as being impossible anyway.
- The property of bounded optimality mentioned at the beginning of this chapter can be paraphrased as "the optimization of computational utility given a set of

[45] See [128] for an axiomatization of qualitative decision making in the style of Savage.

assumptions about expected problems and constraints in reasoning resources" [192]. According to [323], a program exhibits bounded optimality if it "is a solution to the constrained optimization problem presented by its architecture." A relaxation of this concept is *asymptotic* bounded rationality [323] which to some extent parallels the idea of asymptotic complexity. It aims at supporting a constructive theory of bounded rationality which makes the design of bounded optimal agents largely independent of the architecture of the computational environment.

The fact that computational aspects of rational decision making have only recently become a focus of research should not give rise to the impression that related problems have been ignored before. Indeed, classical decision theory has well been aware of computational problems [255]. See, for instance, [182] for a generalization of the axioms of subjective probability taking related aspects into account.

# 8. Conclusions and Outlook

In this book, we have developed various approaches to what we have called *case-based inference*. The idea of CBI is to exploit experience in the form of a memory of observed cases (a case base consisting of input–output tuples) in order to predict a set of promising candidate outputs given a new query input. The corresponding inference schemes are based on suitable formalizations of the heuristic assumption that similar inputs yield similar outputs. Proceeding from a very simple, constraint-based model of this hypothesis, more sophisticated versions have been developed within different formal frameworks of approximate reasoning and reasoning under uncertainty. Let us again highlight the following properties of our approaches:

- For many of the CBI inference schemes, it was possible to derive interesting theoretical properties, for example the fact that a prediction covers the true outcome with high probability. From a case-based reasoning point of view, such CBI methods support a "reliable" retrieval of candidate solutions and, hence, contribute to the formal foundation of an important step within a CBR process.
- As our inference schemes hardly assume more than the specification of similarity measures for inputs and outputs, they are quite general and widely applicable. In particular, since no kind of transitivity is assumed for the similarity measures, the structure of the input and output space might be weaker than that of a metric space. This is a point of great practical relevance for CBR, where inputs and outputs can be complex objects. It also means that predictions can be derived in many situations where standard methods (e.g. from statistics) are not applicable.
- Our inference schemes are applicable for any pair of similarity measures, even if these measures are not defined in an optimal way. That is, the predictions remain correct, even though they might become rather imprecise. This, however, should not be seen as a disadvantage. On the contrary, these methods do not pretend a precision or credibility of case-based predictions which is actually not justified. Instead, imprecise predictions can be taken as an indication that either CBR is not appropriate for the application, or at least that the similarity measures are not well specified.

Most experiments conducted in this book have focused on prediction problems like classification and regression, for which benchmark data is available and predictive accuracy can easily be measured. Of course, from a (case-based) problem solving point of view, prediction appears to be the most simple problem class, mainly because there is no need for adapting the predicted solution. Still, an open question concerns the integration of our CBI methods into more complex CBR systems, that is, the use of these methods for more general types of problem solving.

A interesting idea in this regard is to apply CBI in the context of "search-oriented" CBR. In fact, according to the view of transformational adaptation taken in [30], case-based problem solving can be cast as a search process. Within the related model, (potential) cases correspond to search states and adaptation operators play the role of search operators. Now, the key idea is to use CBI in order to complement this model in a reasonable way. In fact, in [30] the authors note that, according to their approach, CBR could principally be realized by enumerating the search space completely. Understandably, they look at this idea with reservation, immediately pointing to the enormous complexity it brings about. Our approach applies exactly to this problem: CBI supports problem solving by predicting a promising subset of search states (outputs), thereby focusing search to promising regions of the search space and thus providing important information to a search method which is applied for actually finding a solution. From the perspective of CBR, this approach might not merely be seen as an application. In conjunction with the ideas presented in [30], it could contribute in a more general way to a formal framework of CBR in which (transformational) adaptation is realized as a search process and (case-based) experience is used in order to concentrate on promising regions of the related search space.

Indeed, in [30], the concept of similarity is integrated into problem solving by means of a, say, "ideal" similarity measure. By pointing to optimal initial search states, this measure somehow guarantees the retrieval of cases which can be adapted easily. Needless to say, finding such measures will be difficult in practice, if not impossible. As mentioned previously, our CBI methods take a different (more pragmatic) approach: They take any similarity measure as a *given* input, even if this measure is not "ideal", and then derive a *set* of *promising* search states rather than *the optimal* initial state.

# A. Possibilistic Dominance in Qualitative Decisions

Recall the axiomatic system O which has been discussed in the context of qualitative decision making in Section 7.5.3:

O1 $\sqsubseteq$ is a total preorder.

O2 $\pi \leq \pi' \Rightarrow \pi \sqsubseteq \pi'$.

O3 Independence: $\pi_1 \sim \pi_2 \Rightarrow (\lambda/\pi_1, \mu/\pi) \sim (\lambda/\pi_2, \mu/\pi)$.

O4 $\forall \pi \in \mathit{\Pi} \; \exists \lambda \in V \, : \, \pi \sim (\lambda/r^*, 1/r_*)$.

From these axioms one derives the decision criterion (7.48):

$$\mathrm{QU}^+(\pi) \overset{\mathrm{df}}{=} \max_{r \in \mathcal{R}} \min \{h(\pi(r)), u(r)\} \; .$$

The idea of *possibilistic dominance* is the following: Consider the possibility to obtain an outcome which is equal to or better than some fixed outcome $r$. If this possibility is never smaller under a distribution $\pi$ than under a distribution $\pi'$, regardless of the outcome $r$, then $\pi'$ should not be preferred strictly to $\pi$.

**Definition A.1 (possibilistic dominance).** A distribution $\pi \in \mathit{\Pi}$ *dominates* a distribution $\pi' \in \mathit{\Pi}$ *possibilistically* if[1]

$$\forall \, r \in \mathcal{R} \, : \, \pi'(\{x \in \mathcal{R} \, | \, r \sqsubseteq x\}) \leq \pi(\{x \in \mathcal{R} \, | \, r \sqsubseteq x\}) \; .$$

The relation $\sqsubseteq$ satisfies *possibilistic dominance* if $\pi' \sqsubseteq \pi$ whenever $\pi$ dominates $\pi'$ possibilistically. □

If we introduce the *decumulative possibility distribution function* of a distribution $\pi \in \mathit{\Pi}$ via

$$G_\pi : \mathcal{R} \longrightarrow [0,1] \, , \, x \mapsto \pi(\{r \in \mathcal{R} \, | \, x \sqsubseteq r\}) = \max\{\pi(r) \, | \, x \sqsubseteq r\},$$

it can be seen that $\pi$ dominates $\pi'$ possibilistically iff $G_\pi \geq G_{\pi'}$. Note that $\pi \in \mathit{\Pi} \Rightarrow G_\pi \in \mathit{\Pi}$ and that $G_{G_\pi} = G_\pi$ for all $\pi \in \mathit{\Pi}$.

[1] We use the same symbol for a possibility distribution $\pi$ on a (finite) set $X$ and the associated measure which is defined on $2^X$ by $A \mapsto \sup_{x \in A} \pi(x)$.

Remark A.2. Assume (without loss of generality) that $r_1 \sqsubseteq r_2 \sqsubseteq \ldots \sqsubseteq r_n$. Let $\pi \in \mathit{\Pi}$ and define $\pi'$ as follows: $\pi'(r_n) = \pi(r_n)$ and

$$\pi'(r_k) = \max\{\pi(r_k), \pi'(r_{k+1})\}$$

for $k = n-1, n-2, \ldots, 1$. Then $\pi' = G_\pi$. □

**Proposition A.3.** The axiomatic system which consists of O1, O3, O4, and

PD $\sqsubseteq$ satisfies possibilistic dominance

is equivalent to the system O, i.e., O2 can be replaced by PD. □

**Proof.** Suppose that $\pi \leq \pi'$ and that PD holds true. From $\pi \leq \pi'$ follows obviously that $G_\pi \leq G_{\pi'}$, and hence $\pi \sqsubseteq \pi'$. Thus, PD implies O2. We are now going to show that the axiomatic system O implies PD. Again, assume $r_1 \sqsubseteq r_2 \sqsubseteq \ldots \sqsubseteq r_n$. For $1 \leq k \leq n$ define the distributions $\pi_k$ and $\pi_{\leq k}$ as follows:

$$\pi_k(r_j) = \begin{cases} 1 \text{ if } j = k \\ 0 \text{ if } j \neq k \end{cases}, \qquad \pi_{\leq k}(r) = \begin{cases} 1 \text{ if } j \leq k \\ 0 \text{ if } j > k \end{cases}.$$

It is readily seen that O implies $\pi_k \sim \pi_{\leq k}$. Therefore, by axiom O3,

$$\pi' \overset{\text{df}}{=} (\lambda/\pi_{\leq k}, 1/\pi) \sim (\lambda/\pi_k, 1/\pi) \tag{A.1}$$

for all $\pi \in \mathit{\Pi}$. For $\lambda$ in (A.1) equal to $\pi(r_k)$ we obtain

$$(\lambda/\pi_k, 1/\pi) = (\pi(r_k)/\pi_k, 1/\pi) = \pi$$

and

$$\pi'(r_j) = \begin{cases} \max\{\pi(r_j), \pi(r_k)\} & \text{if } j \leq k \\ \pi(r_j) & \text{if } j > k \end{cases}.$$

Now, for $\pi \in \mathit{\Pi}$ let $\pi'_n = \pi$ and

$$\pi'_k = (\pi(r_k)/\pi_{\leq k}, 1/\pi'_{k+1})$$

for $k = n-1, n-2, \ldots, 1$. Then

$$\pi \sim \pi'_n \sim \pi'_{n-1} \sim \ldots \sim \pi'_1 .$$

Moreover, from the construction of $\pi'_1$ and Remark A.2 we obtain $\pi'_1 = G_\pi$. We have thus established that O implies $\pi \sim G_\pi$ for all $\pi \in \mathit{\Pi}$. Now, consider $\pi$, $\pi' \in \mathit{\Pi}$ and suppose that $\pi$ dominates $\pi'$ possibilistically, which means $G_{\pi'} \leq G_\pi$. From axiom O2 we conclude that $G_{\pi'} \sqsubseteq G_\pi$ and, therefore,

$$\pi' \sim G_{\pi'} \sqsubseteq G_\pi \sim \pi .$$

This means that $\pi' \sqsubseteq \pi$ and, hence, that PD holds true. □

Remark A.4. The axiomatic system P can be modified in a similar way with a slightly different definition of possibilistic dominance: A distribution $\pi \in \mathit{\Pi}$ dominates a distribution $\pi' \in \mathit{\Pi}$ possibilistically if

$$\forall\, r \in \mathcal{R} \,:\, \pi(\{x \in \mathcal{R} \,|\, x \sqsubseteq r\}) \leq \pi'(\{x \in \mathcal{R} \,|\, x \sqsubseteq r\}) \ .$$

That is, the possibility of obtaining an outcome which is equal to or worse than a certain fixed outcome is never larger under $\pi$ than under $\pi'$. Again, the pessimistic resp. optimistic character of the decision criteria becomes obvious. According to the axiomatic system O, an optimistic decision maker concentrates on the possibility of receiving a preferable outcome, whereas a pessimistic decision maker, acting in accordance with the axiomatic system P, tries to avoid less preferred outcomes. □

# B. Implication-Based Fuzzy Rules as Randomized Gradual Rules

The probabilistic interpretation of the certainty rule model of case-based inference (cf. Section 6.3) suggests to consider a certainty rule or, more generally, an implication-based fuzzy rule as a class of gradual rules endowed with a probability measure. It has already been mentioned that this kind of representation is unique for certain (implication-based) fuzzy rules. This uniqueness is particularly interesting since a corresponding property does generally not hold in connection with the probabilistic models discussed in Section 4.5. Here, we are going to study the gradual rule representation of implication-based fuzzy rules in more detail. More specifically, this representation is shown to be unique on the assumption that the implication operator used for modeling the fuzzy rule does not have a special kind of strict monotonicity condition. In this case, the crisp relations induced by the involved gradual rules correspond to level-cuts of the fuzzy relation associated with the fuzzy rule. However, other representations might exist if the aforementioned property is not satisfied. Under a slightly stronger (strict) monotonicity condition, the existence of further (non-consonant) representations is even guaranteed. Then, the crisp relations induced by gradual rules do not necessarily correspond to level-cuts of the underlying fuzzy relation.

From a semantical point of view it is often useful to "decompose" a "fuzzy object" into a collection of "crisp objects," i.e., to consider the former as a kind of aggregation of the latter. In fact, tracing the fuzzy case back to the crisp case often supports the understanding of a fuzzy concept and clarifies the meaning of graded degrees of membership (e.g., in terms of possibility, similarity or preference). A clear semantics in turn facilitates the definition and the use of fuzzy concepts, e.g., the determination of membership functions in the context of linguistic modeling, or the learning of fuzzy concepts from observed data.

A well-known example of the above type of decomposition is the interpretation of a fuzzy set in terms of a random set [110, 177, 387]. According to this view, the membership function $A : \mathcal{X} \longrightarrow [0,1]$ of a fuzzy set $A \subset \mathcal{X}$ is considered as the one-point-coverage of a random set, i.e., a random variable $S : \Omega \longrightarrow 2^{\mathcal{X}}$ defined over a probability space $(\Omega, \mathcal{A}, \mu)$. If $\Omega$ is countable, that means

$$A(x) = \sum_{x \in X} \mu(S^{-1}(X)).$$

It deserves mentioning that the representation of a fuzzy set in terms of a random set is in general not unique [177]. However, a decomposition that appears natural is defined by the family of $\alpha$-cuts of a fuzzy set. Let us consider the special case where the fuzzy set is actually a fuzzy relation $C = A \times B$ with membership function $(x, y) \mapsto C(x, y) = \min\{A(x), B(y)\}$. The $\alpha$-cuts of $C$ are then of the form $C_\alpha = A_\alpha \times B_\alpha$. A fuzzy relation of this type is closely connected with so-called conjunction-based fuzzy rules which have originally been used by MAMDANI in the context of fuzzy control [259]. In fact, such a rule actually corresponds to a conjunction rather than an implication, and it is combined disjunctively with other rules (cf. Chapter 5). That is, a fuzzy rule "if $X$ is $A$ then $Y$ is $B$" is formally interpreted as a (fuzzy) logical conjunction $(X \in A) \wedge (Y \in B)$. Thus, looking at the fuzzy relation $A \times B$ associated with a conjunction-based rule as a collection of crisp relations $A_\alpha \times B_\alpha$ comes down to interpreting a fuzzy rule of the Mamdani-type as an aggregation of crisp (Mamdani) rules "if $X$ is $A_\alpha$ then $Y$ is $B_\alpha$."

Here, we are interested in a corresponding representation of implication-based fuzzy rules. Thus, the idea is to represent an implication-based fuzzy rule as a collection of crisp (implication-based) rules. As will be seen, crisp rules can again be associated with $\alpha$-cuts of the fuzzy relation $C$ induced by an implication-based fuzzy rule with antecedent $X \in A$ and consequent $Y \in B$. However, in analogy to the aforementioned non-uniqueness of the representation of a fuzzy set in terms of a random set, other classes of rules can be defined which induce the same fuzzy rule but which do not correspond to a collection of $\alpha$-cuts $C_\alpha$. That is, the decomposition of $C$ into level-cuts is only one possibility of defining a compatible class of crisp rules. In fact, it should be noted that – in the context of implication-based fuzzy rules – this type of decomposition does not appear more "natural" than other decompositions, the associated rules of which do not correspond to level-cuts.

In Section B.1, we introduce implication-based fuzzy rule and discuss (pure) gradual rules as a special case of this type of rule. Even though some of the material has already been presented in Chapter 6, we recall these concepts in order to make this part self-contained. In Section B.2, an interpretation of fuzzy rules in terms of a class of gradual rules endowed with a probability measure is proposed. Section B.3 investigates relations between the uniqueness of this type of representation and monotonicity properties of the implication operator used for modeling the fuzzy rule.

## B.1 Implication-based fuzzy rules

Consider two variables $X$ and $Y$ ranging on domains $D_X$ and $D_Y$, respectively. Moreover, let $A$ and $B$ denote fuzzy subsets of $D_X$ and $D_Y$. These fuzzy sets are characterized by membership functions in the form of $D_X \longrightarrow [0, 1]$ and $D_Y \longrightarrow [0, 1]$ mappings, which we also refer to as $A$ and $B$, respectively. That is, $A(x)$ denotes the degree of membership of $x$ in $A$. For the sake of simplicity we assume the range of $A$ and $B$ to be a finite subset $\mathcal{L} \subset [0,1]$. That is, $A(x)$ and $B(y)$ are elements of $\mathcal{L} = \{\lambda_1, \ldots, \lambda_n\}$, where $0 = \lambda_1 < \lambda_2 < \ldots < \lambda_n = 1$.

Fuzzy rules of the form "if $X$ is $A$ then $Y$ is $B$" can be used for depicting (partial) dependencies between the variables $X$ and $Y$, i.e., for characterizing an underlying relation

$$\varphi \subset D_X \times D_Y \tag{B.1}$$

of *possible* or *admissible* tuples $(x, y)$. The concrete form of the constraint depends on the interpretation of the rule, i.e., on the type of (multiple-valued) implication operator which is used for modeling the rule at a formal level [124].

### B.1.1 Gradual rules

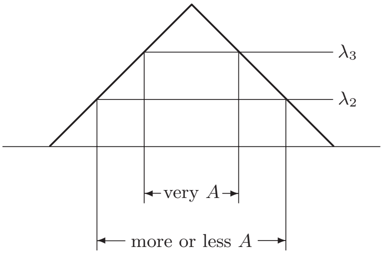


**Fig. B.1.** Linguistic hedges are associated with membership degrees $\lambda \in \mathcal{L}$. Thus, they define level-cuts $A_\lambda$ when being applied to a fuzzy set $A$.

Consider a fuzzy rule like, e.g., "if $X$ is very large then $Y$ is extremely small." This rule can be modeled as a (crisp) constraint of the form

$$X \in A_\alpha \Rightarrow Y \in B_\beta, \tag{B.2}$$

where $A_\alpha = \{x \,|\, A(x) \geq \alpha\}$. $A$ and $B$ represent linguistic labels such as "large" and "small" in our example, and the membership degrees $\alpha, \beta \in \mathcal{L}$ correspond to the related hedges such as "very" and "extremely" (cf. Fig. B.1). The value $\beta$ should be taken as large as possible, so as to make the constraint (B.2) restrictive. On the other hand, $\beta$ has to be defined in such a way that (B.2) is still valid, that is $y \in B_\beta$ for all $(x, y) \in \varphi$ such that $x \in A_\alpha$.

Taking the fuzzy sets $A$ and $B$ as a point of departure, one can thus define a collection

$$X \in A_\lambda \Rightarrow Y \in B_{m(\lambda)} \qquad (\lambda \in \mathcal{L}) \tag{B.3}$$

of constraints (B.2). The function $m : \mathcal{L} \longrightarrow \mathcal{L}$ determines the maximal restriction of $Y$ entailed by conditions of the form $X \in A_\lambda$; such restrictions are expressed in terms of the membership of $Y$ in $B$. One might ask, for instance, to which degree $Y$ is guaranteed to be small if $X$ is *very* large (which means that $X \in A_\lambda$, where $\lambda \in \mathcal{L}$ is associated with the hedge "very"), and the same question can be posed under the condition that $X$ is *more or less* large. Such queries can be answered by an expert in terms of hedges ranging from "not at all" (which corresponds to $m(\lambda) = 0$ in (B.3)) to "completely" ($m(\lambda) = 1$). This way, the expert defines a (linguistic) rule which makes use of the fuzzy sets $A$ and $B$ (see Fig. B.2 for an illustration). Notice that these sets are assumed to be given in advance. This point of view is in line with the *structure-based* approach to rule extraction; it contrasts with alternative (e.g. cluster-based) approaches to learning fuzzy rules where (the membership functions of) $A$ and $B$ are adapted to a predefined rule base [277].

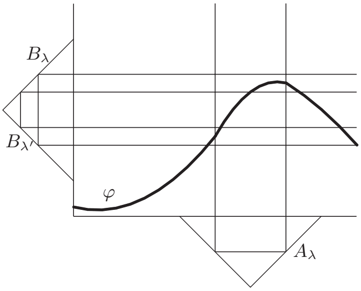


**Fig. B.2.** If $X$ is *more or less* $A$ (i.e., $A(x) \geq \lambda$) then $Y$ is guaranteed to be *more or less* $B$ (i.e. $B(y) \geq \lambda$), but not necessarily *very* $B$. In fact, there are tuples $(x, y)$ in the relation $\varphi$, which here corresponds to a simple function, such that $x \in A_\lambda$ but $y \notin B_{\lambda'}$. For the scale $\mathcal{L} = \{0, \lambda, \lambda', 1\}$ one would hence obtain $m(\lambda) = \lambda$ in (B.3).

The constraints (B.3) can be written compactly in terms of membership functions as

$$m(A(X)) \leq B(Y). \tag{B.4}$$

The inequality (B.4) is often used for expressing the semantics of a so-called *gradual* fuzzy rule (cf. Section 6.1). Such rules, subsequently symbolized by $F \to G$, are of the form "the more $X$ is $F$, the more $Y$ is $G$" or "the larger the degree of membership of $X$ in $F$, the larger the degree of membership of $Y$ in $G$" [119]. In connection with the level cut representation (B.4) we have $F = m \circ A$ and $G = B$ or, alternatively, $F = A$ and $G = m^{(-1)} \circ B$, where $m^{(-1)}$ is defined by

$$m^{(-1)}(\lambda) = \max\{\lambda' \in \mathcal{L} \,|\, m(\lambda') \leq \lambda\}$$

for all $\lambda \in \mathcal{L}$. According to (B.4), a gradual rule $m \circ A \to B$ induces the relation

$$\big\{(x, y) \,|\, m(A(x)) \leq B(y)\big\} \subset D_X \times D_Y$$

of possible or admissible instantiations of $(X, Y)$.

Given two fuzzy sets $A$ and $B$, we can associate a gradual rule $m \circ A \rightarrow B$ with each function $m : \mathcal{L} \longrightarrow \mathcal{L}$. Note that $m$ should be non-decreasing since

$$\big\{y \in D_Y \,|\, \exists\, x \in A_{\lambda'} \,:\, (x, y) \in \varphi\big\} \subset \big\{y \in D_Y \,|\, \exists\, x \in A_{\lambda} \,:\, (x, y) \in \varphi\big\}$$

for $\lambda < \lambda'$, where $\varphi$ is the relation (B.1).[1] Thus, the scale $\mathcal{L}$ gives rise to the class

$$\mathcal{G} = \mathcal{G}_{A,B} = \{m \circ A \rightarrow B \,|\, m(\lambda_0) \leq m(\lambda_1) \leq \ldots \leq m(\lambda_n)\} \tag{B.5}$$

of gradual rules induced by non-decreasing functions $m : \mathcal{L} \longrightarrow \mathcal{L}$.

Observe that $m$ in (B.4) can be used for weakening as well as for strengthening the constraint $A(X) \leq B(Y)$ associated with the "genuine" rule "the more $X$ is $A$, the more $Y$ is $B$." In the extreme case where $m \equiv 0$, for instance, (B.4) is trivially satisfied. Thus, $m \equiv 0$ means that $Y$ cannot be constrained by $X$ in terms of $A$ and $B$. More generally, the function $m$ in (B.4) somehow acts as a *modifier* on the rule $A \rightarrow B$. In connection with the linguistic modeling of fuzzy concepts, modifiers such as $x \mapsto x^2$ or $x \mapsto \sqrt{x}$ are utilized for depicting the effect of linguistic hedges such as "very" or "almost" [244, 413]. It should be noted, however, that such hedges play a slightly different role in our context. Usually, a single linguistic hedge such as "very" is associated with a complete modifier, say, $x \mapsto x^2$. In our approach, a hedge is associated with a membership degree. Therefore, its application to a fuzzy set yields a *level-cut* of that fuzzy set, not a modification thereof which is also a fuzzy set (see again Fig. B.1). (Still one can interpret a level-cut resp. the associated interval as a special fuzzy set. Thus, our approach can be seen as a special case of the general approach, using modifier functions of the form $x \mapsto \mathbb{I}_{[\lambda,1]}$.) None the less, it should be clear that linguistic hedges are only an auxiliary concept which might be interesting from a semantical or practical point of view. Apart from this, they are not needed for the theoretical results discussed below.

### B.1.2 Other implication-based rules

The constraint (B.4) induces a $\{0, 1\}$-valued possibility distribution $\pi_G$ on $D_X \times D_Y$, where $\pi_G(x, y)$ denotes the possibility that $(X, Y) = (x, y)$:

$$\pi_G(x, y) = m(A(x)) \overset{\text{rg}}{\rightsquigarrow} B(y) \tag{B.6}$$

for all $(x, y) \in D_X \times D_Y$, with $\overset{\text{rg}}{\rightsquigarrow}$ being the Rescher-Gaines implication, i.e., $\alpha \overset{\text{rg}}{\rightsquigarrow} \beta = 1$ if $\alpha \leq \beta$ and 0 otherwise. Gradual rules formalized in this way are called *pure gradual rules* in [46]. In fact, a gradual rule $F \rightarrow G$ is a special case of an (implication-based) fuzzy rule of the form $F \rightsquigarrow G$, where $\rightsquigarrow$ is a residuated

[1] This is also in agreement with the "the more ... the more ..." semantics of gradual rules.

(multiple-valued) implication connective. Particularly, a pure gradual rule (which has the same core as any gradual rule $F \rightsquigarrow G$) is obtained by taking $\rightarrow$ as $\overset{\text{rg}}{\rightsquigarrow}$.

More generally, the possibility distribution associated with a rule "if $X$ is $A$ then $Y$ is $B$" is given by

$$\pi_{\rightsquigarrow}(x,y) = A(x) \rightsquigarrow B(y)$$

for all $(x,y) \in \mathcal{L} \times \mathcal{L}$ when making use of a (multiple-valued) implication operator $\rightsquigarrow$. The latter is a $[0,1] \times [0,1] \longrightarrow [0,1]$ function which is non-increasing in the first and non-decreasing in the second argument, i.e.

$$\begin{aligned} \alpha \rightsquigarrow \beta \leq \alpha' \rightsquigarrow \beta \quad &\text{for} \quad \alpha' \leq \alpha, \\ \alpha \rightsquigarrow \beta \leq \alpha \rightsquigarrow \beta' \quad &\text{for} \quad \beta \leq \beta'. \end{aligned} \tag{B.7}$$

Besides, $\rightsquigarrow$ is often assumed to satisfy the following properties [46]:

- identity: $\alpha \rightsquigarrow 1 = 1$,
- exchange: $\alpha \rightsquigarrow (\beta \rightsquigarrow \gamma) = \beta \rightsquigarrow (\alpha \rightsquigarrow \gamma)$,
- neutrality: $1 \rightsquigarrow \beta = \beta$.

Gradual fuzzy rules belong to the class of *truth-qualifying rules*, the semantics of which is adequately modeled by means of so-called R(esiduated)-implications [118]. An R-implication is an operator of the form

$$(\alpha, \beta) \mapsto \sup\{0 \leq \gamma \leq 1 \,|\, \top(\alpha, \gamma) \leq \beta\}, \tag{B.8}$$

where $\top$ is a t-norm (i.e., a function $\top : [0,1] \times [0,1] \longrightarrow [0,1]$ which is associative, commutative, nondecreasing in both arguments, and such that $\top(x,1) = x$ for all $0 \leq x \leq 1$). The Rescher-Gaines implication introduced above is a special case of an R-implication.[2] Other important operators belonging to this class are the Gödel implication and the Goguen implication (as well as the respective contrapositions):

- The Gödel implication

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ \beta & \text{if} \quad \alpha > \beta \end{cases} \tag{B.9}$$

  is the R-implication induced by $\top = \min$.
- The contraposition of the Gödel implication is defined as

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ 1 - \alpha & \text{if} \quad \alpha > \beta \end{cases} . \tag{B.10}$$

[2] It is obtained for the conjunction $\top$ given by $\top(\alpha, \beta) = \alpha$ if $\beta > 0$ and 0 otherwise.

– The Goguen implication

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \alpha = 0 \\ \min\{1, \beta/\alpha\} & \text{if} \quad \alpha > 0 \end{cases} \tag{B.11}$$

is obtained from (B.8) by taking $\top$ as the product $(\alpha, \beta) \mapsto \alpha\beta$.

– The contraposition of the Goguen implication is defined as

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \beta = 1 \\ \min\{1, (1-\alpha)/(1-\beta)\} & \text{if} \quad \beta < 1 \end{cases} . \tag{B.12}$$

A further type of implication is given by so-called S(trong)-implication operators. These are operators of the form $(\alpha, \beta) \mapsto S(n(\alpha), \beta)$, with $S$ and $n$ being respectively a t-conorm and a strong negation function. Such operators adequately capture the semantics of *(un)certainty-qualifying rules* [118]. A *certainty rule* $A \rightsquigarrow B$ corresponds to statements of the form "the more $X$ is $A$, the more *certain* $Y$ lies in $B$." More precisely, it can be interpreted as a collection of rules "if $X = x$, it is certain at least to the degree $A(x)$ that $Y$ lies in $B$." This translates into the constraint $\pi(x, y) \le \max\{1 - A(x), B(y)\}$ when taking $\rightsquigarrow$ as the Kleene-Dienes implication

$$\alpha \rightsquigarrow \beta = \max\{1 - \alpha, \beta\}. \tag{B.13}$$

A further example of an S-implication is the Reichenbach implication

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} 1 - \alpha + \alpha\beta. \tag{B.14}$$

There are also implication operators which belong to both the class of R-implications and the class of S-implications:

– The Lukasiewicz implication

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \min\{1, 1 - \alpha + \beta\} \tag{B.15}$$

is the R-implication induced by the t-norm $(\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$ and defines the S-implication for the t-conorm $(\alpha, \beta) \mapsto \min\{1, \alpha + \beta\}$.

– The operator

$$\alpha \rightsquigarrow \beta \overset{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \alpha \le \beta \\ \max\{1 - \alpha, \beta\} & \text{if} \quad \alpha > \beta \end{cases} \tag{B.16}$$

is the $S$-implication and, at the same time, the $R$-implication related to a t-norm called the nilpotent minimum (given a strong negation $n$, the latter is defined as $\top(\alpha, \beta) = \min\{\alpha, \beta\}$ if $\beta > n(\alpha)$ and $\top(\alpha, \beta) = 0$ otherwise [150]).

Since we consider implication operators defined on a finite scale $\mathcal{L}$, each such operator $\rightsquigarrow$ is identified by an $n \times n$-matrix $\gamma = (\gamma_{\imath\jmath})_{1 \le \imath, \jmath \le n}$, where

$$\gamma_{\imath\jmath} = \lambda_\imath \rightsquigarrow \lambda_\jmath. \tag{B.17}$$

Given a fuzzy rule “If $X$ is $A$ then $Y$ is $B$,” the knowledge of the set of values (B.17) is also sufficient for comparing the effect of different operators as constraints on the relation $\varphi$ in (B.1). In fact, $\gamma$ determines the induced possibility distribution $\pi_\rightsquigarrow$ on $D_X \times D_Y$ completely: The possibility $\pi_\rightsquigarrow(x, y)$ assigned to a tuple $(x, y)$ is given by $\gamma_{\imath\jmath}$, where $\imath$ and $\jmath$ are such that $A(x) = \lambda_\imath$ and $B(y) = \lambda_\jmath$.

## B.2 Randomized gradual rules

Our idea is to establish a relationship between an implication-based rule $A \rightsquigarrow B$ and the class (B.4) of (pure) gradual rules[3] associated with $A$ and $B$. As already mentioned above, the latter can be seen as a class of modifications $m \circ A \rightarrow B$ or $A \rightarrow m^{(-1)} \circ B$ of the genuine (gradual) rule with antecedent $A$ and consequent $B$. Of course, (B.4) is not the only class of crisp rules which might be associated with the fuzzy rule $A \rightsquigarrow B$. Yet, it emerges quite naturally from the constraint-based view expressed by (B.2) when defining the condition and conclusion parts of the rules in terms of of the level-cuts of the fuzzy sets $A$ and $B$. The constraint-based view is in turn natural if rules are considered as implications.

In accordance with the random set interpretation of fuzzy sets we are now going to represent fuzzy rules as *randomized* gradual rules.

**Definition B.1 (randomized gradual rule).** A randomized gradual rule (or random rule for short) associated with a conditional statement “if $X$ is $A$ then $Y$ is $B$” is a tuple $(\mathcal{G}, p)$, where $\mathcal{G} = \mathcal{G}_{A,B}$ is the (finite) set of gradual rules (B.5), and $p$ is a probability distribution on $\mathcal{G}$. Each rule $m \circ A \rightarrow B$ is identified by the corresponding modifier $m : \mathcal{L} \longrightarrow \mathcal{L}$. Moreover, $p(m) = p(m \circ A \rightarrow B)$ is interpreted as the probability of the rule $m \circ A \rightarrow B$. □

Since each gradual rule $m \circ A \rightarrow B$ induces an admissible set

$$\begin{aligned} \varGamma_m &= \big\{(x, y) \in D_X \times D_Y \,|\, m(A(x)) \leq B(y)\big\} \qquad \text{(B.18)} \\ &= \bigcap_{\lambda \in \mathcal{L}} (A_\lambda \times B_{m(\lambda)}) \cup (\overline{A}_\lambda \times D_Y) \end{aligned}$$

of tuples $(x, y)$, a randomized gradual rule $(\mathcal{G}, p)$ gives rise to a random set over $D_X \times D_Y$ such that

$$\mathbb{P}(\varGamma) = \sum_{m \in \mathcal{G} \,:\, \varGamma = \varGamma_m} p(m)$$

for all $\varGamma \subset D_X \times D_Y$. This random set (which is not necessarily nested) defines an upper probability on $D_X \times D_Y$ according to

[3] We shall henceforth use the term *gradual rule* as a synomym for *pure gradual rule*, i.e., we always have the Rescher-Gaines formalization in mind when speaking of a gradual rule.

$$\overline{p}(x,y) = \sum_{\mathit{\Gamma}\,:\,(x,y)\in\mathit{\Gamma}} \mathbb{P}(\mathit{\Gamma}).$$

A random rule $(\mathcal{G},p)$ hence induces a possibility distribution $\pi_{(\mathcal{G},p)}$ on $D_X \times D_Y$, where possibility degrees are interpreted as upper probabilities:

$$\pi_{(\mathcal{G},p)}(x,y) = \overline{p}(x,y) \tag{B.19}$$

for all $(x,y) \in D_X \times D_Y$. Observe that

$$\pi_{(\mathcal{G},p)} = \sum_{m\in\mathcal{G}} p(m)\cdot\pi_m, \tag{B.20}$$

where $\pi_m$ denotes the $\{0,1\}$-valued possibility distribution associated with the gradual rule $m \circ A \rightarrow B$. Moreover, (B.19) is completely determined by the following implication operator associated with $(\mathcal{G},p)$:

$$\lambda_\imath \overset{(\mathcal{G},p)}{\rightsquigarrow} \lambda_\jmath = \gamma_{\imath\jmath} = \sum_{m\in\mathcal{G}\,:\,m(\lambda_\imath)\le\lambda_\jmath} p(m). \tag{B.21}$$

In fact, we have $\pi_{(\mathcal{G},p)}(x,y) = \overline{p}(x,y) = \gamma_{\imath\jmath}$ whenever $A(x) = \lambda_\imath$ and $B(y) = \lambda_\jmath$.

Remark B.2. Equation (B.21) shows that $\gamma_{\imath\jmath} < \gamma_{\imath,\jmath+1}$ as soon as $m(\lambda_\imath) = \jmath + 1$ for a modifier $m$ such that $p(m) > 0$. In other words, $\gamma_{\imath\jmath} = \gamma_{\imath,\jmath+1}$ means that $m(\lambda_\imath) = \jmath + 1$ is impossible. □

Remark B.3. From a mathematical point of view, (B.20) is nothing else than a convex combination of the implications $m \circ A \rightarrow B$. Interpreting the weights $p(m)$ as probabilities is of course not compulsory. □

Remark B.4. The implication $\overset{(\mathcal{G},p)}{\rightsquigarrow}$ defined by (B.21) satisfies (B.7) as well as the properties of identity and exchange, but not necessarily neutrality. □

Example B.5. Let $\mathcal{L} = \{0, 1/2, 1\}$, and consider two gradual rules defined by the modifiers $m_1$ and $m_2$ such that

$$\begin{aligned} &m_1(0) = 0, \quad m_1(1/2) = 1/2, \quad m_1(1) = 1,\\ &m_2(0) = 0, \quad m_2(1/2) = 0, \qquad m_2(1) = 1/2. \end{aligned}$$

These rules induce (crisp) implication operators $\overset{\mathrm{rg}}{\rightsquigarrow}_1$ and $\overset{\mathrm{rg}}{\rightsquigarrow}_2$, respectively:

$$\begin{array}{c|ccc} \overset{\mathrm{rg}}{\rightsquigarrow}_1 & 0 & 1/2 & 1 \\ \hline 0 & 1 & 1 & 1 \\ 1/2 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{array} \qquad \begin{array}{c|ccc} \overset{\mathrm{rg}}{\rightsquigarrow}_2 & 0 & 1/2 & 1 \\ \hline 0 & 1 & 1 & 1 \\ 1/2 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{array} \tag{B.22}$$

According to (B.21), the random rule $(\mathcal{G}, p)$ with $p(m_1) = p(m_2) = 1/2$ defines the following implication operator:

| $\overset{(\mathcal{G},p)}{\rightsquigarrow}$ | 0 | 1/2 | 1 |
|---|---|---|---|
| 0 | 1 | 1 | 1 |
| 1/2 | 1/2 | 1 | 1 |
| 1 | 0 | 1/2 | 1 |

(B.23)

Note that (B.23) corresponds to the operator induced by (B.16) on $\mathcal{L} \times \mathcal{L}$. □

## B.3 A probabilistic representation of implication-based fuzzy rules

In this section, we shall establish the following results: For each implication operator $\rightsquigarrow$, a probability $p$ exists such that the rule $A \rightsquigarrow B$ is equivalent to the random gradual rule $(\mathcal{G}, p)$ in the sense that $\pi_{\rightsquigarrow} = \pi_{(\mathcal{G},p)}$, where the possibility distribution $\pi_{\rightsquigarrow}$ is obtained from $A \rightsquigarrow B$ and $\pi_{(\mathcal{G},p)}$ is given by (B.20). That is, the rule $A \rightsquigarrow B$ and the randomized gradual rule $(\mathcal{G}, p)$ induce the same possibility distribution on $D_X \times D_Y$. Moreover, it is shown that the probability $p$ is guaranteed to be unique (resp. non-unique) under certain conditions on the implication operator $\rightsquigarrow$. Henceforth, we suppose an implication operator to satisfy the monotonicity conditions (B.7) and the condition of identity.

Before turning to the question of the uniqueness of a representation we show the existence of an equivalent random rule.

**Lemma B.6.** For each rule $A \rightsquigarrow B$ an equivalent random gradual rule $(\mathcal{G}, p)$ exists. That is, there is a probability $p$ such that $\pi_{\rightsquigarrow} = \pi_{(\mathcal{G},p)}$, where $\pi_{\rightsquigarrow}$ is the possibility distribution induced by $A \rightsquigarrow B$ and $\pi_{(\mathcal{G},p)}$ denotes the distribution (B.20). □

**Proof.** We prove this lemma by constructing a randomized gradual rule which is equivalent to $A \rightsquigarrow B$. Denote by $\gamma_1 > \gamma_2 > \ldots > \gamma_K > 0$ the elements of the set

$$\mathcal{L}' = \{\lambda \rightsquigarrow \lambda' \,|\, \lambda, \lambda' \in \mathcal{L}\} \setminus \{0\}.$$

For $1 \leq \imath \leq n$ and $1 \leq k \leq K$ let

$$\lambda(\imath, k) = \min\{\lambda_\jmath \,|\, 1 \leq \jmath \leq n,\ \lambda_\imath \rightsquigarrow \lambda_\jmath \geq \gamma_k\}. \tag{B.24}$$

Observe that the set on the right-hand side in (B.24) is not empty due to the identity property and, hence, that (B.24) is well-defined. (Besides, the mapping $(\lambda_\imath, \gamma_k) \mapsto \lambda(\imath, k)$ defines a conjunctive operation $\top$ whose residuation just yields the implication $\rightsquigarrow$.[4] In fact, we have $\lambda_\imath \rightsquigarrow \lambda_\jmath \geq \gamma_k \Leftrightarrow \lambda(\imath, k) = \top(\lambda_\imath, \gamma_k) \leq \lambda_\jmath$.)

[4] Note, however, that this mapping is only defined on $\mathcal{L}' \times \mathcal{L}$.

Now, let the modifier $m_k$ $(1 \leq k \leq K)$ be defined by

$$\forall\, 1 \leq \imath \leq n \,:\, m_k(\lambda_\imath) = \lambda(\imath, k). \tag{B.25}$$

Moreover, let $p(m_k) = \gamma_k - \gamma_{k+1}$ (and $p(m') = 0$ for all $m' \notin \{m_1, \ldots, m_K\}$), where $\gamma_{K+1} = 0$. Observe that $m_k$ thus defined is non-decreasing since the implication $\rightsquigarrow$ is non-increasing in the first argument. It hence determines a gradual rule according to (B.5). Moreover, the probability distribution $p$ defines a random gradual rule $(\mathcal{G}, p)$ in the sense of Definition B.1.

Now, consider any $1 \leq \imath, \jmath \leq n$ such that $\lambda_\imath \rightsquigarrow \lambda_\jmath = \gamma_l$. We obviously have $m_k(\lambda_\imath) \leq \lambda_\jmath \Leftrightarrow \gamma_l \geq \gamma_k$. Thus,

$$\begin{aligned} \gamma_{\imath\jmath} &= \sum_{m \in \mathcal{G} \,:\, m(\lambda_\imath) \leq \lambda_\jmath} p(m) \\ &= (\gamma_l - \gamma_{l+1}) + (\gamma_{l+1} - \gamma_{l+2}) + \ldots + (\gamma_K - \gamma_{K+1}) \\ &= \gamma_l \end{aligned}$$

for the value $\gamma_{\imath\jmath} = \lambda_\imath \overset{(\mathcal{G},p)}{\rightsquigarrow} \lambda_\jmath$ according to (B.21). This means that $\overset{(\mathcal{G},p)}{\rightsquigarrow}$ is equivalent to $\rightsquigarrow$ and, hence, that $(\mathcal{G}, p)$ induces the same possibility distribution on $D_X \times D_Y$ as the rule $A \rightsquigarrow B$. □

We shall call $m \circ A \rightarrow B$ a *focal rule* if $p(m) > 0$. Moreover, we denote by $\mathcal{P}_C$ the class of probability distributions $p$ which induce a *consonant* random rule, where $(\mathcal{G}, p)$ is called consonant if the sets

$$\Lambda_m = \{(\lambda_\imath, \lambda_\jmath) \,|\, m(\lambda_\imath) \leq \lambda_\jmath\} \tag{B.26}$$

associated with focal gradual rules can be arranged into a chain. That is, all pairs of focal rules $m \circ A \rightarrow B$ and $m' \circ A \rightarrow B$ are nested:

$$\Lambda_m \subset \Lambda_{m'} \quad \text{or} \quad \Lambda_{m'} \subset \Lambda_m. \tag{B.27}$$

The focal rules of a consonant random rule can obviously be arranged according to their restrictiveness: The rule associated with $m$ is more restrictive than the one associated with $m'$ if $\Lambda_m \subset \Lambda_{m'}$, i.e., if the admissibility of a tuple $(x, y)$ according to the latter implies the admissibility of $(x, y)$ according to the former.

**Lemma B.7.** Each implication operator has exactly one representation in terms of a consonant randomized gradual rule. □

**Proof.** Consider the randomized gradual rule $(\mathcal{G}, p)$ constructed in Lemma B.6. From (B.24) and (B.25) follows that $m_k \leq m_l$ and, hence, $\Lambda_{m_l} \subset \Lambda_{m_k}$ for $l < k$. Thus, (B.27) is satisfied, i.e., the probability $p$ is an element of $\mathcal{P}_C$. This proves the existence of a consonant representation. Uniqueness follows from the unique representation of a possibility distribution (which here corresponds to $(\lambda_\imath, \lambda_\jmath) \mapsto \gamma_{\imath\jmath}$) in terms of a consonant body of evidence [232]. □

The existence of a unique consonant random rule representing an implication operator (which satisfies identity) does not exclude the existence of other representations which are non-consonant. The following lemma gives a sufficient condition for an implication to have a unique (and hence consonant) representation in terms of a randomized gradual rule.

**Lemma B.8.** Let $(\mathcal{G}, p)$ be a randomized gradual rule and suppose focal rules $m_1, m_2 \in \mathcal{G}$ and associated sets $\varLambda_{m_1}, \varLambda_{m_2} \subset \boldsymbol{\mathcal{L}} \times \boldsymbol{\mathcal{L}}$ to exist such that neither $\varLambda_{m_1} \subset \varLambda_{m_2}$ nor $\varLambda_{m_2} \subset \varLambda_{m_1}$. Then, the property

$$\exists\, \imath < k \;\exists\, \jmath < l \;:\; (\gamma_{k\jmath} < \gamma_{\imath\jmath} < \gamma_{\imath l}) \;\wedge\; (\gamma_{k\jmath} < \gamma_{kl} < \gamma_{\imath l}). \tag{B.28}$$

is satisfied. □

**Proof.** Let $(\lambda_\imath, \lambda_\jmath) \in \varLambda_{m_1} \setminus \varLambda_{m_2}$ and $(\lambda_k, \lambda_l) \in \varLambda_{m_2} \setminus \varLambda_{m_1}$, i.e.

$$\begin{aligned} m_1(\lambda_\imath) \le \lambda_\jmath, \quad m_1(\lambda_k) > \lambda_l, \\ m_2(\lambda_\imath) > \lambda_\jmath, \quad m_2(\lambda_k) \le \lambda_l. \end{aligned}$$

Assume without loss of generality that $\imath \le k$. We first show that $\imath < k$ and $\jmath < l$. Indeed, $\imath = k$ and $\jmath \le l$ leads to the contradiction $m_1(\lambda_\imath) \le \lambda_\jmath \le \lambda_l < m_1(\lambda_k) = m_1(\lambda_\imath)$. Likewise, $\imath = k$ and $\jmath > l$ yields $m_2(\lambda_\imath) > \lambda_\jmath > \lambda_l \ge m_2(\lambda_k) = m_2(\lambda_\imath)$. Therefore, $\imath < k$ must hold. Now, assume $\jmath \ge l$. We then obtain $m_2(\lambda_k) \le \lambda_l \le \lambda_\jmath < m_2(\lambda_\imath)$ which is impossible since $m_2$ is non-decreasing. Thus, $\imath < k$ and $\jmath < l$.

Let us now show the inequalities in (B.28). From $\imath < k$, $\jmath < l$ and the monotonicity property of $m_2$ follows

$$\begin{aligned} m_1(\lambda_\imath) \le \lambda_\jmath < \lambda_l, \quad m_2(\lambda_\imath) \le m_2(\lambda_k) \le \lambda_l, \\ m_1(\lambda_k) > \lambda_l > \lambda_\jmath, \quad m_2(\lambda_k) \ge m_2(\lambda_\imath) > \lambda_\jmath. \end{aligned}$$

That means

$$(\lambda_\imath, \lambda_l) \in \varLambda_{m_1} \cap \varLambda_{m_2} \quad \text{and} \quad (\lambda_k, \lambda_\jmath) \notin \varLambda_{m_1} \cup \varLambda_{m_2}.$$

Hence, both probabilities $p(m_1)$ and $p(m_2)$ appear in the computation of $\gamma_{\imath l}$ according to (B.21), while $p(m_2)$ is not assigned to $\gamma_{\imath\jmath}$ nor $p(m_1)$ to $\gamma_{kl}$. Moreover, neither $p(m_1)$ nor $p(m_2)$ is assigned to $\gamma_{k\jmath}$. Therefore,

$$\begin{aligned} \gamma_{\imath l} - \gamma_{\imath\jmath} &\ge p(m_2) > 0, \\ \gamma_{\imath l} - \gamma_{kl} &\ge p(m_1) > 0, \\ \gamma_{\imath\jmath} - \gamma_{k\jmath} &\ge p(m_1) > 0, \\ \gamma_{kl} - \gamma_{k\jmath} &\ge p(m_2) > 0, \end{aligned}$$

which completes the proof. □

Remark B.9. The condition (B.28) can be illustrated by arranging the values (B.21) in the form of a table. In fact, (B.28) is satisfied if this table contains two rows and two columns such that the four entries at the respective intersections have a unique minimum (the bottom left element) and a unique maximum (the top right element):

$$\begin{array}{c|ccccccc}
 & \lambda_1 & \dots & \lambda_\jmath & \dots & \lambda_l & \dots & \lambda_n \\
\hline
\lambda_1 & & & & & & & \\
\vdots & & & & & & & \\
\lambda_\imath & & & \gamma_{\imath\jmath} & < & \gamma_{\imath l} & & \\
\vdots & & & > & & > & & \\
\lambda_k & & & \gamma_{k\jmath} & < & \gamma_{kl} & & \\
\vdots & & & & & & & \\
\lambda_n & & & & & & &
\end{array}$$

As can be seen, (B.28) means that the implication $\rightsquigarrow$ is *strictly monotone* in both places on a subset of $\mathcal{L} \times \mathcal{L}$ (namely on $\{\lambda_\imath, \lambda_k\} \times \{\lambda_\jmath, \lambda_l\}$). □

**Theorem B.10.** For each rule $A \rightsquigarrow B$ formalized by means of an implication operator $\rightsquigarrow$ an equivalent (consonant) random rule $(\mathcal{G}, p)$ exists. Moreover, $(\mathcal{G}, p)$ is a unique random rule representation if

$$\forall\, \imath < k \,\forall\, \jmath < l \,:\, \neg(\gamma_{k\jmath} < \gamma_{\imath\jmath} < \gamma_{\imath l}) \;\vee\; \neg(\gamma_{k\jmath} < \gamma_{kl} < \gamma_{\imath l}), \tag{B.29}$$

i.e., if $\rightsquigarrow$ does not satisfy the monotonicity condition (B.28). □

**Proof.** Existence has already been shown in Lemma B.6. The uniqueness in the case where $\rightsquigarrow$ satisfies (B.29) follows from Lemma B.7 and Lemma B.8. □

Let us summarize the results obtained so far. We have shown that any fuzzy rule $A \rightsquigarrow B$, where $\rightsquigarrow$ is an implication operator (i.e., non-increasing in the first and non-decreasing in the second argument) satisfying the identity property, can be represented as a randomized gradual rule. One of these representations is always consonant. Moreover, this consonant representation is the only way of expressing $A \rightsquigarrow B$ in terms of a randomized gradual rule if the (non-)monotonicity condition (B.29) holds.

**Corollary B.11.** A rule $A \rightsquigarrow B$ with $\rightsquigarrow$ the Kleene-Dienes implication (B.13) can be expressed uniquely in terms of an equivalent (consonant) random rule. □

**Proof.** According to Theorem B.10 we only have to show that (B.29) is satisfied (i.e., that (B.28) is not satisfied) for the implication operator $\alpha \rightsquigarrow \beta = \max\{1 - \alpha, \beta\}$: We have $\gamma_{\imath l} = \max\{1 - m(\lambda_\imath), \lambda_l\}$ and $\gamma_{kl} = \max\{1 - m(\lambda_k), \lambda_l\}$. Now, suppose that (B.28) holds. Then, inequality $\gamma_{kl} < \gamma_{\imath l}$ implies $\lambda_l \leq \gamma_{kl} < \gamma_{\imath l}$ and, hence,

$$\gamma_{\imath l} = 1 - m(\lambda_\imath) \leq \max\{1 - m(\lambda_\imath), \lambda_\jmath\} = \gamma_{\imath\jmath},$$

which is in contradiction to $\gamma_{\imath\jmath} < \gamma_{\imath l}$. □

**Corollary B.12.** A rule $A \rightsquigarrow B$ with $\rightsquigarrow$ being the Gödel implication (B.9) or its contraposition (B.10) can be expressed uniquely in terms of an equivalent (consonant) random rule. □

**Proof.** Again, we have to show that (B.29) is satisfied for (B.10): Suppose $\gamma_{k\jmath} < \gamma_{kl}$. Since $\gamma_{k\jmath}, \gamma_{kl} \in \{1 - \lambda_k, 1\}$, we then have $\gamma_{kl} = 1$ and, hence, $\gamma_{\imath l} = \gamma_{k\jmath}$, which means that (B.29) holds. Now, suppose that $\gamma_{kl} < \gamma_{\imath l}$. That means $\gamma_{kl} = 1 - \lambda_k$ (since otherwise $\gamma_{kl} = 1 \geq \gamma_{\imath l}$) and, hence, $\gamma_{k\jmath} = 1 - \lambda_k = \gamma_{kl}$. Thus, it is not possible to satisfy both, $\lambda_{k\jmath} < \lambda_{kl}$ and $\lambda_{kl} < \lambda_{\imath l}$. The result for the operator (B.9) is shown in the same way. □

The construction in Lemma B.6 shows that the (unique) consonant random rule $(\mathcal{G}, p)$ is a "levelwise" reconstruction of the underlying rule $A \rightsquigarrow B$. That is, for each level $\alpha \in \mathcal{L}'$ there is a focal rule with associated modifier $m_\alpha$. This gradual rule corresponds to the $\alpha$-cut of the fuzzy rule $A \rightsquigarrow B$ in the sense that

$$\{(x, y) \,|\, A(x) \rightsquigarrow B(y) \geq \alpha\} = \{(x, y) \,|\, m_\alpha(A(x)) \leq B(y)\}, \tag{B.30}$$

where $m_\alpha(\beta) = \top(\alpha, \beta)$ and $\top$ is the conjunctive operation whose residuation yields the implication $\rightsquigarrow$. (Note that $m_\alpha(A)$ is not necessarily normalized. In this respect, the gradual rule $m_\alpha(A) \rightarrow B$ is somewhat different from the gradual rules considered in [119, 124] which involve only normal fuzzy sets.) If the representation of $A \rightsquigarrow B$ is not unique, however, further equivalent random rules exist, and these rules are not consonant. That is, they do not correspond to a levelwise representation of the form (B.30).

Remark B.13. Note that the $\alpha$-cuts of the fuzzy relation $C$ induced by a fuzzy rule $A \rightsquigarrow B$ can generally not be expressed as a function of the $\alpha$-cuts of $A$ and $B$, as it is the case for Mamdani rules. For instance, $C_\alpha$ does generally not correspond to $(\overline{A}_\alpha \times D_Y) \cup (A_\alpha \times B_\alpha)$, where $D_Y$ is the domain of $Y$ and $\overline{A}_\alpha = D_X \setminus A_\alpha$ denotes the complement of $A_\alpha$. □

Example B.14. Consider as an example the implication $\rightsquigarrow$ in (B.10) for $\mathcal{L} = \{0, 1/2, 1\}$ and identify a modifier $m : \mathcal{L} \rightarrow \mathcal{L}$ with the vector

$$(m(0), m(1/2), m(1)) \in \mathcal{L}^3.$$

The fuzzy rule $A \rightsquigarrow B$ is then equivalent to $(\mathcal{G}, p)$, where

$$p((0, 1/2, 1)) = 1/2, \quad p((0, 0, 1)) = 1/2.$$

Observe that the set of constraints

$$\begin{array}{llll} \text{If} & A(x) = 0 & \text{then} & B(y) \geq 0, \\ \text{If} & A(x) = 1/2 & \text{then} & B(y) \geq 1/2, \\ \text{If} & A(x) = 1 & \text{then} & B(y) = 1, \end{array} \tag{B.31}$$

can be associated with the gradual rule $(0, 1/2, 1)$. Likewise, the rule $(0, 0, 1)$ gives rise to

$$\begin{array}{llll} \text{If} & A(x) = 0 & \text{then} & B(y) \geq 0, \\ \text{If} & A(x) = 1/2 & \text{then} & B(y) \geq 0, \\ \text{If} & A(x) = 1 & \text{then} & B(y) = 1. \end{array} \tag{B.32}$$

Obviously, (B.31) and (B.32) only differ with respect to the conclusion drawn from $A(x) = 1/2$. The randomized gradual rule $(\mathcal{G}, p)$ can thus be interpreted as follows: If the premise $X \in A$ is not satisfied at all, nothing can be said about the value of $Y$ (as expressed by the trivial constraints $B(y) \geq 0$ in (B.31) and (B.32)). As opposed to this, the conclusion $Y \in B$ is completely satisfied whenever the same is true for the premise. However, the rule that $Y$ is (at least) more or less (i.e., to the degree 1/2) in $B$ if $X$ is more or less in $A$ is valid only with probability 1/2. □

Example B.15. Interestingly enough, the uniqueness results shown in Theorem B.10 does not apply to the implication operator (B.16): For $\mathcal{L} = \{0, 1/2, 1\}$, the fuzzy rule based on (B.16) is equivalent to the randomized gradual rules $(\mathcal{G}, p)$ and $(\mathcal{G}, p')$, where

$$\begin{array}{ll} p((0, 0, 1/2)) = 1/2, & p((0, 1/2, 1)) = 1/2, \\ p'((0, 0, 1)) = 1/2, & p'((0, 1/2, 1/2)) = 1/2. \end{array}$$

Observe that $(\mathcal{G}, p')$ is not consonant since neither $m_1 \leq m_2$ nor $m_2 \leq m_1$ holds true for $m_1 = (0, 0, 1)$ and $m_2 = (0, 1/2, 1/2)$. That is, neither is $m_1$ more restrictive than $m_2$ nor vice versa. □

Remark B.16. Simple counterexamples (using scales $\mathcal{L}$ which contain at least two elements different from 0 and 1) can also be constructed in order to show that the uniqueness in Theorem B.10 does not apply to the following operators: The Goguen implication (B.11) and its contraposition (B.12), the Reichenbach implication (B.14), and the Lukasiewicz implication (B.15). These examples again reflect the fact that a fuzzy set (which here corresponds to the fuzzy relation $(\lambda, \lambda') \mapsto \lambda \rightsquigarrow \lambda'$) does generally not have a unique representation in terms of a random set (which is given here in the form of a randomized gradual rule). □

Theorem B.10 has shown that the (non-monotonicity) property (B.29) is a sufficient condition for the representation of a fuzzy rule in terms of a randomized gradual rule to be unique. It is, however, not a necessary condition. In other words, property (B.28) does not imply the existence of a non-consonant random rule representation. This can be illustrated by means of the following implication operator which satisfies (B.28) (with $\imath = 2, \jmath = 1, k = 4, l = 3$), but which can only be represented in terms of a consonant random rule:[5]

| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ |
|---|---|---|---|---|
| $\lambda_1$ | 1 | 1 | 1 | 1 |
| $\lambda_2$ | 1/2 | 1/2 | 1 | 1 |
| $\lambda_3$ | 1/2 | 1/2 | 1/2 | 1 |
| $\lambda_4$ | 0 | 1/2 | 1/2 | 1 |

Even though (B.28) is not sufficient for the existence of a non-consonant representation, it can be shown that a slightly stronger condition actually is. This condition requires an implication operator $\rightsquigarrow$ to be strictly monotone (in both places) on an interval of the form $\{\lambda_a, \lambda_{a+1}\} \times \{\lambda_b, \lambda_{b+1}\} \subset \mathcal{L} \times \mathcal{L}$.[6] That is, there are values $1 \leq a, b < n$ such that

$$\big(\gamma_{a+1,b} < \gamma_{ab} < \gamma_{a,b+1}\big) \;\wedge\; \big(\gamma_{a+1,b} < \gamma_{a+1,b+1} < \gamma_{a,b+1}\big), \tag{B.33}$$

where $\gamma$ denotes the matrix (B.17) induced by $\rightsquigarrow$. In order to prove this result we need some preliminaries.

**Definition B.17 (admissible fuzzy relation).** We call a fuzzy relation $\gamma \subset \mathcal{L} \times \mathcal{L}$ identified by the matrix $(\gamma_{\imath\jmath})_{1 \leq \imath,\jmath \leq n}$ *admissible* if it satisfies the weak monotonicity property

$$(\imath \leq k) \wedge (\jmath \leq l) \;\Rightarrow\; (\gamma_{k\jmath} \leq \gamma_{\imath l}) \tag{B.34}$$

for all $1 \leq \imath, \jmath, k, l \leq n$ and if

$$\gamma_{\imath n} = \gamma_{\jmath n} \tag{B.35}$$

for all $1 \leq \imath, \jmath \leq n$. Note that an admissible relation is not assumed to be normal. That is, $\max_{1 \leq \imath,\jmath \leq n} \gamma_{\imath\jmath} < 1$ is not excluded. □

**Definition B.18 (partial random rule).** Suppose two fuzzy sets $A \subset D_X$ and $B \subset D_Y$ to be given. A partial random rule is a tuple $(\mathcal{G}, p)$, where $\mathcal{G} = \mathcal{G}_{A,B}$ is the (finite) set of gradual rules (B.5), and $p$ is a mapping on $\mathcal{G}$ such that $0 \leq p(m) = p(m \circ A \rightarrow B) \leq 1$ for all $m \in \mathcal{G}$ and $\sum_{m \in \mathcal{G}} p(m) \leq 1$. □

**Lemma B.19.** For each admissible fuzzy relation $\gamma$ there is an equivalent partial random rule $(\mathcal{G}, p)$, i.e., a rule $(\mathcal{G}, p)$ such that

$$\gamma_{\imath\jmath} = \sum_{m \in \mathcal{G} \,:\, m(\lambda_\imath) \leq \lambda_\jmath} p(m) \tag{B.36}$$

for all $1 \leq \imath, \jmath \leq n$. Moreover, each fuzzy relation induced by a partial random rule via (B.36) is admissible. □

[5] This follows immediately from Remark B.2.

[6] In other words, the two rows as well as the two columns in the table in Remark B.9 are directly neighbored.

**Proof.** The first part can be shown by means of a slight adaptation of the proof of Lemma B.6. The second part is obvious. □

**Lemma B.20.** Denote by $\gamma_{\rightsquigarrow} \subset \mathcal{L} \times \mathcal{L}$ the fuzzy relation (B.17) induced by the implication operator $\rightsquigarrow$. Let $(\mathcal{G}, p')$ be a partial random rule and let $\gamma'$ be the induced (admissible) fuzzy relation (B.36). Moreover, suppose that $\gamma' \leq \gamma_{\rightsquigarrow}$ and that $\gamma_{\rightsquigarrow} - \gamma'$ is admissible. There is a partial random rule $(\mathcal{G}, p'')$ such that $(G, p' + p'')$ is a randomized gradual rule which is equivalent to $A \rightsquigarrow B$. □

**Proof.** Since $\gamma'' = \gamma_{\rightsquigarrow} - \gamma'$ is admissible, Lemma B.19 guarantees the existence of an equivalent partial random rule $(\mathcal{G}, p'')$. Let $p = p' + p''$. It is obvious that $(\mathcal{G}, p)$ defines a partial random rule with induced fuzzy relation $\gamma = \gamma' + \gamma'' = \gamma_{\rightsquigarrow}$. Now, recall that $\rightsquigarrow$ is assumed to satisfy the identity property, which means that

$$\max_{1 \leq \imath, \jmath \leq n} \gamma_{\imath\jmath} = \max_{1 \leq \imath, \jmath \leq n} \gamma_{\rightsquigarrow}(\lambda_\imath, \lambda_\jmath) = 1.$$

Moreover, note that $m(\lambda_1) \leq \lambda_n$ for all $m \in \mathcal{G}$. Therefore,

$$\sum_{m \in \mathcal{G}} p(m) = \gamma_{1n} = 1,$$

i.e., $(\mathcal{G}, p)$ is a randomized gradual rule. □

**Theorem B.21.** Denote by $\gamma = \gamma_{\rightsquigarrow} \subset \mathcal{L} \times \mathcal{L}$ the matrix (B.17) induced by the implication operator $\rightsquigarrow$ and suppose that $\gamma$ satisfies the strict monotonicity condition (B.33). There is a non-consonant randomized gradual rule $(\mathcal{G}, p)$ which is equivalent to $\rightsquigarrow$. □

**Proof.** In order to prove the theorem we proceed as follows:

(i) We define two functions $m_1, m_2 : \mathcal{L} \longrightarrow \mathcal{L}$.

(ii) Both functions are shown to be monotone, i.e., $m_1$ and $m_2$ are elements of $\mathcal{G}$ in (B.5).

(iii) It is verified that the sets $\Lambda_1$ and $\Lambda_2$ defined, respectively, by $m_1$ and $m_2$ according to (B.26) are not nested. Therefore, a partial random rule $(\mathcal{G}, p')$ such that $p'(m_1) > 0$ and $p'(m_2) > 0$ is not consonant.

(iv) It is shown that positive values $p'(m_1)$ and $p'(m_2)$ can be defined such that $\gamma - \gamma'$ is admissible, where $\gamma$ and $\gamma'$ are the relations induced by $\rightsquigarrow$ and $(\mathcal{G}, p')$, respectively.

(v) According to Lemma B.20, $(\mathcal{G}, p')$ can be extended to a randomized gradual rule $(\mathcal{G}, p)$ which is equivalent to $\rightsquigarrow$. Since $(\mathcal{G}, p')$ is non-consonant, so is $(\mathcal{G}, p)$. This concludes the proof.

(i) Let $1 \leq a, b < n$ be given such that (B.33) is satisfied. Note that we have $b < n-1$ due to the identity property of $\rightsquigarrow$. We define $m_1 : \mathcal{L} \longrightarrow \mathcal{L}$ as follows: $m_1(\lambda_\imath) = \lambda_{b_\imath}$ for all $1 \leq \imath \leq n$, where

$$b_\imath = \begin{cases} \min\{\jmath \,|\, \gamma_{\imath\jmath} \geq \gamma_{ab}\} & \text{if} \quad \imath \leq a \\ \min\{\jmath \,|\, \gamma_{\imath\jmath} > \gamma_{a+1,b+1}\} & \text{if} \quad \imath > a \end{cases} .$$

Moreover, $m_2 : \mathcal{L} \longrightarrow \mathcal{L}$ is given by $m_2(\lambda_\imath) = \lambda_{b_\imath}$ for all $1 \leq \imath \leq n$, where

$$b_\imath = \begin{cases} \min\{\jmath \,|\, \gamma_{\imath\jmath} \leq \gamma_{ab}\} & \text{if} \quad \imath \leq a \\ \min\{\jmath \,|\, \gamma_{\imath\jmath} > \gamma_{a+1,b}\} & \text{if} \quad \imath > a \end{cases} . \tag{B.37}$$

(ii) The monotonicity of $m_1$ on $\{1, \ldots, a\}$ and $\{a+1, \ldots, n\}$ follows immediately from the monotonicity property (B.34) of the relation $\gamma$ induced by $\rightsquigarrow$. Moreover,

$$m_1(\lambda_a) \leq \lambda_b < \lambda_{b+1} < m_1(\lambda_{a+1})$$

due to (B.33). Thus, $m_1$ is indeed monotone. The monotonicity of $m_2$ is shown in the same way.

(iii) Denote by $\Lambda_{m_1}$ and $\Lambda_{m_2}$ the sets induced, respectively, by $m_1$ and $m_2$ according to (B.26). From the definition of $m_1$ and $m_2$ follows that $(\lambda_a, \lambda_b) \in \Lambda_1 \setminus \Lambda_2$ and $(\lambda_{a+1}, \lambda_{b+1}) \in \Lambda_2 \setminus \Lambda_1$. Thus, $\Lambda_{m_1}$ and $\Lambda_{m_2}$ are not nested:

$$\Lambda_{m_1} \not\subset \Lambda_{m_2} \quad \text{and} \quad \Lambda_{m_2} \not\subset \Lambda_{m_1} .$$

(iv) First, we verify that the following property holds true:

$$(\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1} \wedge (\lambda_k, \lambda_l) \in \Lambda_{m_1} \Rightarrow \gamma_{\imath\jmath} < \gamma_{kl} \tag{B.38}$$

for all $1 \leq \imath, \jmath, k, l \leq n$. In view of (B.34) and the transitivity of $\leq$, (B.38) is obviously equivalent to

$$\begin{aligned} (\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1} \wedge (\lambda_{\imath-1}, \lambda_\jmath) \in \Lambda_{m_1} \Rightarrow \gamma_{\imath\jmath} < \gamma_{\imath-1,\jmath}, \\ (\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1} \wedge (\lambda_\imath, \lambda_{\jmath+1}) \in \Lambda_{m_1} \Rightarrow \gamma_{\imath\jmath} < \gamma_{\imath,\jmath+1}. \end{aligned} \tag{B.39}$$

Suppose $(\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1}$ and notice that $\jmath < n$ due to the identity property of $\rightsquigarrow$. Moreover, the definition of $m_1$ obviously implies that (B.39) is satisfied if $\imath = 1$. Thus, let $1 < \imath \leq n$. We distinguish the following cases:

– $(\lambda_{\imath-1}, \lambda_\jmath) \in \Lambda_{m_1}$ and $(\lambda_\imath, \lambda_{\jmath+1}) \notin \Lambda_{m_1}$.
If $\imath \leq a$ this means that $\gamma_{\imath\jmath} < \gamma_{ab} \leq \gamma_{\imath-1,\jmath}$. Likewise, if $\imath > a+1$ it means that $\gamma_{\imath\jmath} \leq \gamma_{a+1,b+1} < \gamma_{\imath-1,\jmath}$. Thus, (B.39) is satisfied in both situations. For the case $\imath = a+1$ we can make use of (B.33): If $\jmath \leq b-1$ then

$$\gamma_{\imath\jmath} \leq \gamma_{a+1,b} < \gamma_{ab} \leq \gamma_{\imath-1,\jmath}.$$

If $\jmath = b$ then

$$\gamma_{\imath\jmath} = \gamma_{a+1,b} < \gamma_{ab} = \gamma_{\imath-1,b}.$$

Finally, if $\jmath \geq b+1$ then

$$\gamma_{\imath\jmath} \leq \gamma_{a+1,b+1} < \gamma_{a,b+1} \leq \gamma_{\imath-1,\jmath}.$$

- $(\lambda_{\imath-1}, \lambda_\jmath) \notin \Lambda_{m_1}$ and $(\lambda_\imath, \lambda_{\jmath+1}) \in \Lambda_{m_1}$.
  If $\imath \leq a$ this means $\gamma_{\imath\jmath} < \gamma_{ab} \leq \gamma_{\imath,\jmath+1}$. Likewise, if $\imath > a+1$ then $\gamma_{\imath\jmath} \leq \gamma_{a+1,b+1} < \gamma_{\imath,\jmath+1}$. The case $\imath = a+1$ is actually not possible. In fact, $(\lambda_\imath, \lambda_{\jmath+1}) \in \Lambda_{m_1}$ implies $\gamma_{a+1,\jmath+1} > \gamma_{a+1,b+1}$, i.e., $\jmath \geq b+1$. Then, however, $\gamma_{\imath-1,\jmath} \geq \gamma_{a,b+1} > \gamma_{ab}$, which contradicts $(\lambda_{\imath-1}, \lambda_\jmath) \notin \Lambda_{m_1}$.
- $(\lambda_{\imath-1}, \lambda_\jmath) \in \Lambda_{m_1}$ and $(\lambda_\imath, \lambda_{\jmath+1}) \in \Lambda_{m_1}$.
  If $\imath \leq a$ this means that $\gamma_{\imath\jmath} < \gamma_{ab} \leq \gamma_{\imath-1,\jmath}$ and $\gamma_{\imath\jmath} < \gamma_{ab} \leq \gamma_{\imath,\jmath+1}$. If $\imath > a+1$ then $\gamma_{\imath\jmath} \leq \gamma_{a+1,b+1} < \gamma_{\imath-1,\jmath}$ and $\gamma_{\imath\jmath} \leq \gamma_{a+1,b+1} < \gamma_{\imath,\jmath+1}$. Now, let $\imath = a+1$. Again, $\gamma_{\imath\jmath} \leq \gamma_{a+1,b+1} < \gamma_{\imath,\jmath+1}$. Moreover, $(\lambda_\imath, \lambda_{\jmath+1}) \in \Lambda_{m_1}$ implies $\jmath \geq b+1$. Therefore,
  $$\gamma_{\imath\jmath} < \gamma_{a+1,b+1} < \gamma_{a,b+1} \leq \gamma_{a\jmath} = \gamma_{\imath-1,\jmath}.$$

Since property (B.38) holds true, we can find some number $p'(m_1) > 0$ such that the relation $\gamma'$ defined by

$$\gamma'_{\imath\jmath} = \begin{cases} \gamma_{\imath\jmath} & \text{if} \quad (\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1} \\ \gamma_{\imath\jmath} - p'(m_1) & \text{if} \quad (\lambda_\imath, \lambda_\jmath) \in \Lambda_{m_1} \end{cases}$$

satisfies the same monotonicity properties as the relation $\gamma$. That is, $p'(m_1)$ can be chosen small enough such that

$$\gamma_{\imath\jmath} \leq \gamma_{kl} \Leftrightarrow \gamma'_{\imath\jmath} \leq \gamma'_{kl} \quad \text{and} \quad \gamma_{\imath\jmath} < \gamma_{kl} \Leftrightarrow \gamma'_{\imath\jmath} < \gamma'_{kl} \tag{B.40}$$

for all $1 \leq k \leq \imath \leq n$ and $1 \leq \jmath \leq l \leq n$. In particular, $\gamma'$ still satisfies the monotonicity property (B.34), i.e., it is still admissible. Indeed, (B.40) is obvious if the tuples $(\imath, \jmath)$ and $(k, l)$ are such that $(\lambda_\imath, \lambda_\jmath) \in \Lambda_{m_1}$ and $(\lambda_k, \lambda_l) \in \Lambda_{m_1}$ or $(\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1}$ and $(\lambda_k, \lambda_l) \notin \Lambda_{m_1}$. Otherwise, if $(\lambda_\imath, \lambda_\jmath) \notin \Lambda_{m_1}$ and $(\lambda_k, \lambda_l) \in \Lambda_{m_1}$, we can find a sequence of elements $\gamma_{u_m,v_m}$ $(m = 1, \ldots, M)$ such that

- $(u_1, v_1) = (\imath, \jmath)$, $(u_M, v_M) = (k, l)$,
- $(u_{m+1}, v_{m+1}) = (u_m - 1, v_m)$ or $(u_{m+1}, v_{m+1}) = (u_m, v_m + 1)$,
- $\gamma_{u_m,v_m} \leq \gamma_{u_{m+1},v_{m+1}}$ for all $1 \leq m < M$,
- there is exactly one $1 \leq m < M$ such that
  $$(\lambda_{u_m}, \lambda_{v_m}) \notin \Lambda_{m_1}, \quad (\lambda_{u_{m+1}}, \lambda_{v_{m+1}}) \in \Lambda_{m_1}$$
  and, hence, $\gamma_{u_m,v_m} < \gamma_{u_{m+1},v_{m+1}}$.

By choosing $p'(m_1)$ small enough, the strict inequality in the last point remains unchanged, i.e., $\gamma'_{\imath\jmath} = \gamma_{\imath\jmath} < \gamma_{kl} - p'(m_1) = \gamma'_{kl}$.

By making use of (B.40), it can be shown in the same way as above that (B.38) holds true with $\gamma$ and $\mathit{\Lambda}_{m_1}$ replaced by $\gamma'$ and $\mathit{\Lambda}_{m_2}$, respectively. Thus, there is again a number $p'(m_2) > 0$ such the relation $\gamma''$ given by

$$\gamma''_{\imath\jmath} = \begin{cases} \gamma_{\imath\jmath} & \text{if} \quad (\lambda_\imath, \lambda_\jmath) \notin \mathit{\Lambda}_{m_1} \cup \mathit{\Lambda}_{m_2} \\ \gamma_{\imath\jmath} - p'(m_1) & \text{if} \quad (\lambda_\imath, \lambda_\jmath) \in \mathit{\Lambda}_{m_1} \setminus \mathit{\Lambda}_{m_2} \\ \gamma_{\imath\jmath} - p'(m_2) & \text{if} \quad (\lambda_\imath, \lambda_\jmath) \in \mathit{\Lambda}_{m_2} \setminus \mathit{\Lambda}_{m_1} \\ \gamma_{\imath\jmath} - p'(m_1) - p'(m_2) & \text{if} \quad (\lambda_\imath, \lambda_\jmath) \in \mathit{\Lambda}_{m_1} \cap \mathit{\Lambda}_{m_2} \end{cases}$$

is still monotone in the sense of (B.34). Moreover, property (B.35) holds since $(\lambda_\imath, \lambda_n) \in \mathit{\Lambda}_{m_1} \cap \mathit{\Lambda}_{m_2}$ for all $1 \leq \imath \leq n$, i.e., $\gamma''$ is admissible. Therefore, the partial random rule $(\mathcal{G}, p')$ defined by the positive numbers $p'(m_1), p'(m_2)$ (and $p'(m) = 0$ for $m \notin \{m_1, m_2\}$) has the property mentioned in (iv) above.

The argument given in (v) hence concludes the proof. □

# C. Similarity-Based Reasoning as Logical Inference

The methods of similarity-based (case-based) inference presented in previous chapters make use of special types of models in order to formalize the CBI hypothesis, and thereby combine model-based and instance-based reasoning (cf. Section 2.1). Each type of model especially supports a certain aspect of similarity-based reasoning, such as the consideration of uncertainty or the incorporation of background knowledge into the inference process. Moreover, our approaches are *data-driven* in the sense that the models can be used for adapting similarity-based inference to the application at hand. An alternative, more logic-oriented formalization of similarity-based reasoning was proposed by Ruspini [320] and has later been pursued by other authors as well (e.g. [102]). In the following we briefly sketch the basic ideas underlying this approach.

Suppose a reflexive, symmetric, and $\top$-transitive similarity relation $\sigma$ to be defined over the interpretations $U$ associated with a formal language $\mathcal{L}$, where $\top$ is a t-norm (cf. Section 2.3). One might then be tempted to look for inference rules which take the similarity between interpretations (i.e., assignments of truth values for each propositional variable in $\mathcal{L}$) into account. For instance, one could think of generalizing the classical modus ponens as follows ($\alpha$ and $\beta$ are propositions in $\mathcal{L}$):

$$\frac{\begin{array}{c}\alpha \text{ is close to being true}\\ \alpha \text{ approximately implies } \beta\end{array}}{\beta \text{ is not far from being true}}$$

For an interpretation $u$ and a proposition $p$ we write $u \models p$ if $u$ is a model of $p$. The set of models of $p$ is denoted $[p]$. For a subset $A \subset U$, the fuzzy set

$$A^* : u \mapsto \sup_{v \in A} \sigma(u, v)$$

is the fuzzy set of elements which are close to $A$. Thus, the proposition "approximately $p$" can be identified with the fuzzy set of models $[p^*] \overset{\text{df}}{=} [p]^*$, where $p$ is an ordinary proposition. Based on this, one can introduce a *graded satisfaction relation* $\models^\alpha$ as follows:

$$u \models^\alpha p \quad \text{iff} \quad \exists\, v \in U \,:\, v \models p \wedge u \in [v^*]_\alpha,$$

where $[v^*]_\alpha$ is the $\alpha$-cut of the fuzzy set $[v^*]$.

A *graded semantic entailment relation* can be derived from $\models^\alpha$ as follows: A proposition $p$ entails a proposition $q$ at the level $\alpha$ if each model of $p$ makes $q$ at least $\alpha$-true:

$$p \models^\alpha q \quad \text{iff} \quad [p] \subset [q^*]_\alpha .$$

The relation $\models^\alpha$ satisfies

| | |
|---|---|
| nestedness: | $p \models^\alpha q$ and $\beta \leq \alpha$ implies $p \models^\beta q$, |
| $\top$-transitivity: | $p \models^\alpha r$ and $r \models^\beta q$ implies $p \models^{\top(\alpha,\beta)} q$, |
| reflexivity: | $\forall\, \alpha \in [0,1] \,:\, p \models^\alpha q$, |
| right weakening: | $p \models^\alpha q$ and $q \models r$ implies $p \models^\alpha r$, |
| left strengthening: | $p \models r$ and $r \models^\alpha q$ implies $r \models^\alpha q$, |
| left OR: | $p \wedge q \models^\alpha r$ iff $p \models^\alpha r$ and $q \models^\alpha r$. |

Is is worth mentioning that $\models^\alpha$ does not satisfy the “right and” property. That is, $p \models^\alpha q$ and $p \models^\alpha r$ does not imply $p \models^\alpha q \wedge r$. Therefore, the class of (approximate) consequences of $p$, in the sense of $\models^\alpha$, is not deductively closed.

See [410] and [34] for yet another type of similarity logic, where inference is approximate in the sense that the antecedent clause of a rule is allowed to match its premise only approximately. This approach is syntactical in nature, whereas the one outlined above is a semantical one.

# D. Simulation Results of Section 3.4.4

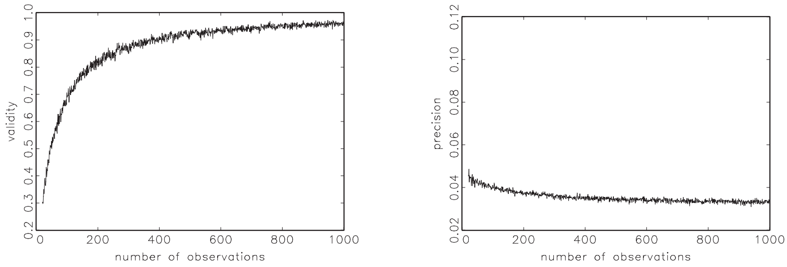


**Fig. D.1.** Correctness and precision for the case-based approximation of the function $\boldsymbol{s} \rightarrow \sin(s+1)$ $\cos^2(s)$, using (3.15).

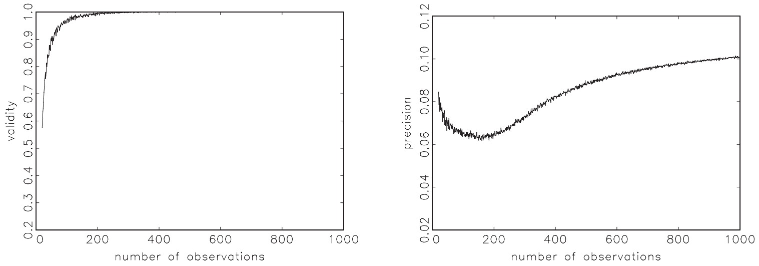


**Fig. D.2.** Correctness and precision for the case-based approximation of the function $s \mapsto \sin(s+1)$ $\cos^2(s)$, using (3.16) with $k = 10$.

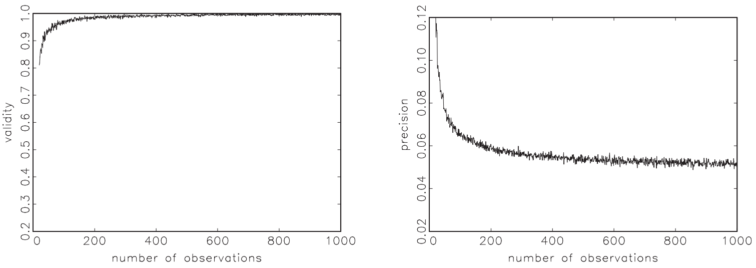


**Fig. D.3.** Correctness and precision for the case-based approximation of the function $s \mapsto \sin(s+1) \cdot \cos^2(s)$, using (3.16) with 10% of the observed cases.

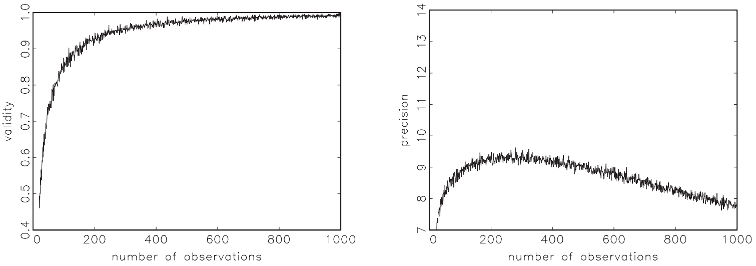


**Fig. D.4.** Correctness and precision for the case-based approximation of the function $\varphi$ defined for the CBI setup $\Sigma_1$, using (3.15).

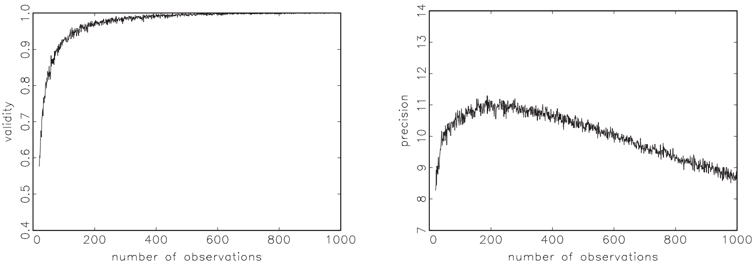


**Fig. D.5.** Correctness and precision for the case-based approximation of the function $\varphi$ defined for the CBI setup $\Sigma_1$, using (3.16) with $k = 10$.

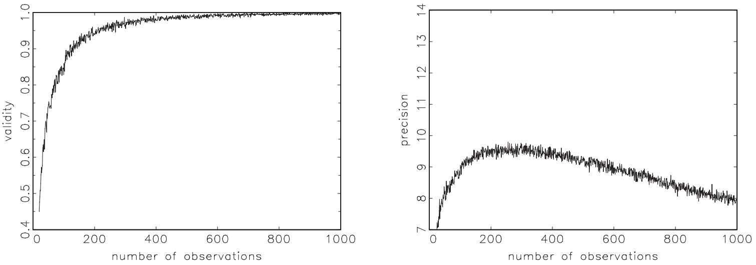


**Fig. D.6.** Correctness and precision for the case-based approximation of the function $\varphi$ defined for the CBI setup $\Sigma_1$, using (3.16) with $k = 100$.

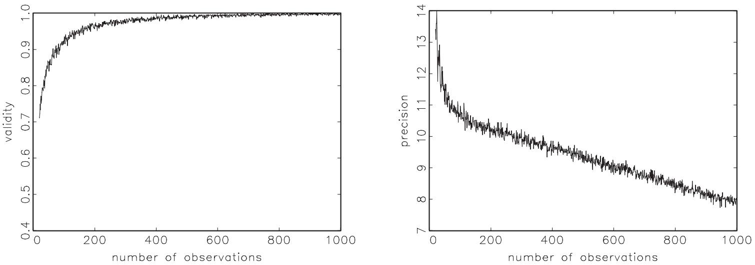
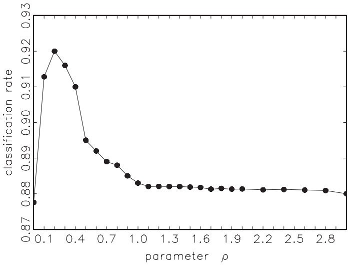
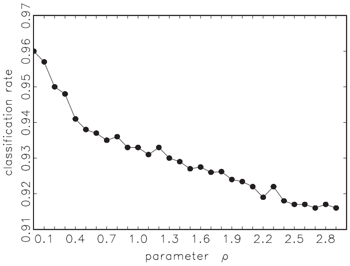


**Fig. D.7.** Correctness and precision for the case-based approximation of the function $\varphi$ defined for the CBI setup $\Sigma_1$, using (3.16) with 10% of the observed cases.

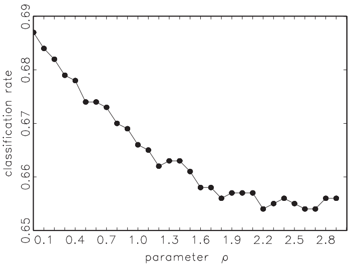
# E. Experimental Results of Section 5.5.4



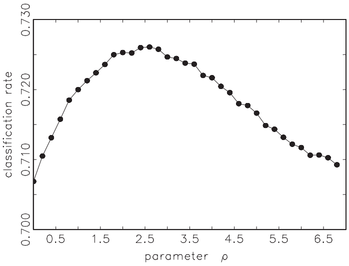
Experimental results for the Balance Scale data.



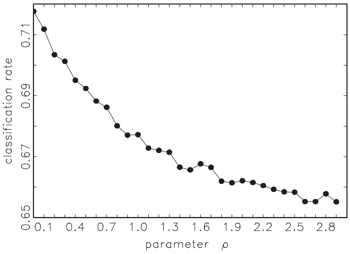
Experimental results for the Iris Plant data.



Experimental results for the Glass Identification data.



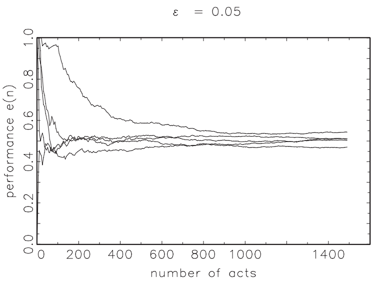
Experimental results for the Pima Indian Diabetes data.

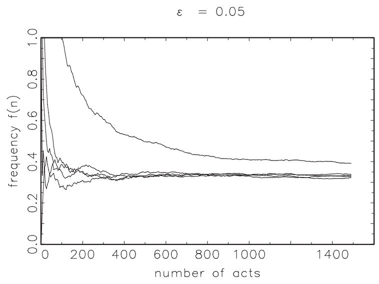


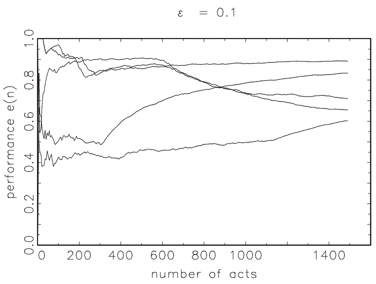
Experimental results for the Wine Recognition data.

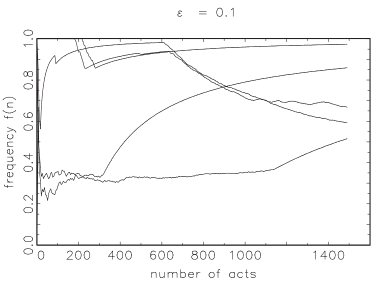
# F. Simulation Results of Section 7.4

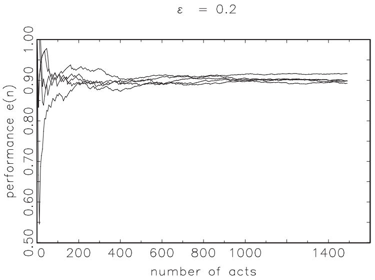
The pictures on the left plot the performance $e(n)$ of a decision strategy for several simulation runs, where $e(n) = \sum_{k=1}^{n} u(k)/n$ and $u(k) \in U = \{0, 1\}$ denotes the outcome for the $k$-th decision. The pictures on the right plot the respective frequencies $f(n) = \sum_{k=1}^{n} c(k)/n$ of "correct" decisions, i.e., $c(k) = 1$ if the $k$-th decision was $a_1$ and $c(k) = 0$ otherwise.

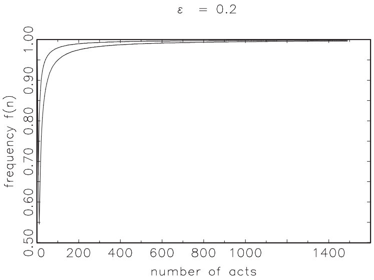


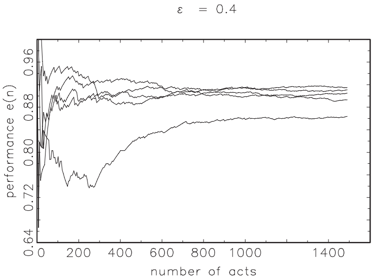


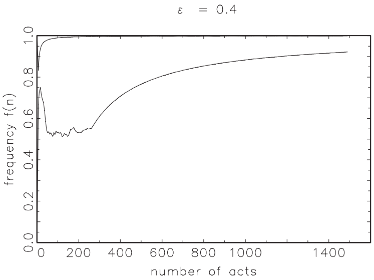


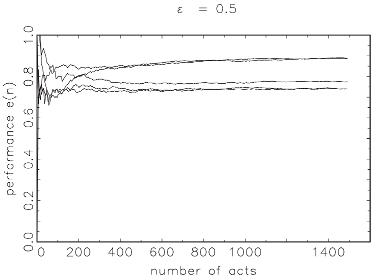


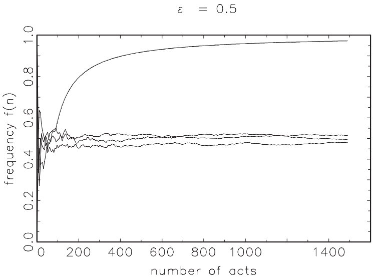

ε = 0.2


ε = 0.2


ε = 0.4


ε = 0.4


ε = 0.5


ε = 0.5

# G. Computation of an Extended Splitting Measures

A possible approach to computing the extended splitting measure (7.63) in Section 7.7.2 is to search the space of selections of $S^*$ in a systematic way. This can be realized very efficiently by means of a branch & bound procedure when using (7.59) as an impurity measure. In fact, a good bounding function can be derived on the basis of the following result.

**Proposition G.1.** Consider subsets $A_1, \ldots, A_n$ of a finite set $\{1, \ldots, N\}$. For each selection $a = (a_1, \ldots, a_n) \in A_1 \times \ldots \times A_n$ define the probability vector $p(a) = (p_1, \ldots, p_n)$ such that $p_\imath = N_\imath / N$, where $N_\imath = \mathrm{card}\{1 \leq \jmath \leq N \,|\, a_\jmath = \imath\}$. Now, let $(a_1, \ldots, a_m)$ be a partial selection of the first $m < n$ values and denote by $\mathcal{C}$ the class of possible completions $a = (a_1, \ldots, a_n)$. Then the following holds ($p \cdot p$ denotes the dot product $p_1 \, p_1 + \ldots + p_N \, p_N$ of a vector $p$):

$$\max_{a \in \mathcal{C}} p(a) \cdot p(a) \leq p(a^c) \cdot p(a^c),$$

where the vector $a^c$ (which is not necessarily a feasible selection!) is defined as

$$a^c_\imath = \begin{cases} a_\imath & \text{if} \quad 1 \leq \imath \leq m \\ a_{max} & \text{if} \quad m+1 \leq \imath \leq n \end{cases}$$

and $1 \leq a_{max} \leq N$ is any number (e.g., the smallest) such that

$$\mathrm{card}\{1 \leq \imath \leq m \,|\, a_\imath = a_{max}\} \geq \mathrm{card}\{1 \leq \imath \leq m \,|\, a_\imath = k\}$$

for all $1 \leq k \leq N$. □

**Proof.** Without loss of generality we can assume $a_{max} = 1$. Let $k = n - m$ and $m_\jmath = \mathrm{card}\{1 \leq \imath \leq m \,|\, a_\imath = \jmath\}$ for $1 \leq \jmath \leq N$. The probability vector $p(a^c)$ is then given by

$$\left( \frac{m_1 + k}{n}, \frac{m_2}{n}, \ldots, \frac{m_N}{n} \right),$$

and any other completion $a$ of $(a_1, \ldots, a_m)$ yields the vector

$$\left( \frac{m_1 + k_1}{n}, \frac{m_2 + k_2}{n}, \ldots, \frac{m_N + k_N}{n} \right),$$

where $k_1 + \ldots + k_N = k$. Thus, we have

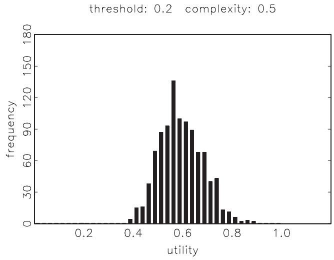
$$
\begin{aligned}
n^2 \left(p(a^c) \cdot p(a^c) - p(a) \cdot p(a)\right) &= \\
&= (m_1 + k)^2 + \sum_{\imath=2}^{N} (m_\imath)^2 - \sum_{\imath=1}^{N} (m_\imath + k_\imath)^2 \\
&= \sum_{\imath=1}^{N} (m_\imath)^2 + 2\, m_1\, k + k^2 - \sum_{\imath=1}^{N} (m_\imath)^2 - \\
&\qquad\qquad - 2 \sum_{\imath=1}^{N} m_\imath\, k_\imath - \sum_{\imath=1}^{N} (k_\imath)^2 \\
&= 2 \underbrace{\left( m_1\, k - \sum_{\imath=1}^{N} m_\imath\, k_\imath \right)}_{\geq 0} + \underbrace{\left( k^2 - \sum_{\imath=1}^{N} (k_\imath)^2 \right)}_{\geq 0} \geq 0,
\end{aligned}
$$

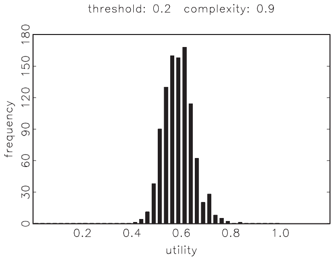
which means $p(a) \cdot p(a) \leq p(a^c) \cdot p(a^c)$. □

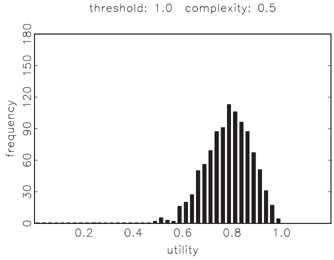
The above proposition suggests an effective lower bound to the GINI function (7.59) and, hence, an upper bound to the information gain (7.58). In fact, this upper bound can be used as a bounding function in a branch & bound algorithm which works as follows: Suppose a set $S^*$ of generalized examples (7.60) and an attribute $T$ to be given. In order to compute $m(T, S^*)$, a selection of the set $A_{p_m}$ of actions is chosen at the $m$th level of the algorithm. That is, each node of the branch & bound tree at level $m$ has $\text{card}(A_{p_m})$ successors, each of which corresponds to the choice of one particular act $a_m \in A_{p_m}$. At each node, the aforementioned bounding function can be derived for the associated partial selection $(a_1, \ldots, a_m)$. A branch (partial selection) is continued only if the value of the bounding function is above the current best solution (information gain). Moreover, the following heuristic strategies turned out to be useful in practice: The sets $A_{p_\imath}$ in the extended sample (7.60) should be re-arranged according to their size (smaller sets before larger ones). Moreover, the order in which actions $a_\imath \in A_{p_\imath}$ are chosen (i.e., the order of successors of an inner node of the search tree) should reflect the following preference: An act $a$ is preferred to (chosen before) an act $a'$ if it is more frequent among the sets $A_{p_\imath}$, that is, $\text{card}\{\imath \,|\, a \in A_{p_\imath}\} \geq \text{card}\{\imath \,|\, a' \in A_{p_\imath}\}$.

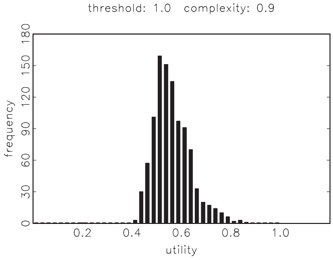
# H. Experimental Results of Section 7.7.2

The following pictures show the empirical distributions (histograms) of the average utility degrees for different combinations of $u^*$ and $\gamma$:

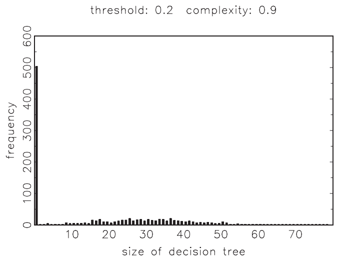


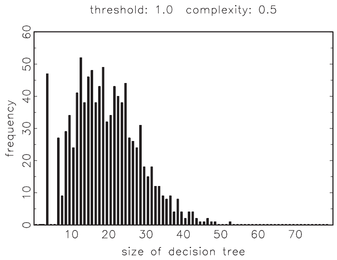


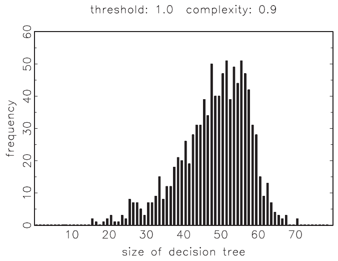




The following pictures show the empirical distributions of the average number of leaf nodes for different combinations of $u^*$ and $\gamma$:


threshold: 0.2 complexity: 0.9
frequency
0
100
200
300
400
500
600
10
20
30
40
50
60
70
size of decision tree


threshold: 1.0 complexity: 0.5
frequency
0
10
20
30
40
50
60
10
20
30
40
50
60
70
size of decision tree


threshold: 1.0 complexity: 0.9
frequency
0
10
20
30
40
50
60
10
20
30
40
50
60
70
size of decision tree

# References

1. A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

2. S. Abe and M.S. Lan. Fuzzy rules extraction directly from numerical data for function approximation. IEEE *Transactions on Systems, Man, and Cybernetics*, 25(1):119–129, 1995.

3. D.W. Aha. Incremental, instance-based learning of independent and graded concept descriptions. In *Proceedings of the 6th International Workshop on Machine Learning*, pages 387–391, Ithaca, NY, 1989. Morgan Kaufmann.

4. D.W. Aha. Case-based learning algorithms. In R. Bareiss, editor, *Proceedings of the* DAPRA *Workshop on Case-Based Reasoning*, pages 147–158. Morgan Kaufmann Publishers, 1991.

5. D.W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36:267–287, 1992.

6. D.W. Aha, editor. *Lazy Learning*. Kluwer Academic Publ., 1997.

7. D.W. Aha and R.L. Bankert. A comparative evaluation of sequential feature selection algorithms. In D. Fisher and J.H. Lenz, editors, *Artificial Intelligence and Statistics*, Ney York, 1996. Springer-Verlag.

8. D.W. Aha and L.W. Chang. Cooperative Bayesian and case-based reasoning for solving multiagent planning tasks. Technical Report AIC-96-005, Navy Center for Applied Research in AI, Naval Research Laboratory, Washington, D.C., U.S.A., 1996.

9. D.W. Aha and R.L. Goldstone. Concept learning and flexible weighting. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 534–539, Bloominghton, IN, 1992. Lawrence Erlbaum.

10. D.W. Aha and D. Kibler. Noise-tolerant instance-based learning algorithms. In *Proceedings* IJCAI-89*, 11th International Joint Conference on Artificial Intelligence*, pages 794–799, Detroit, 1989. Morgan Kaufmann.

11. D.W. Aha, D. Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

12. D.W. Aha and D. Wettscherek. Case-based learning: Beyond classification of feature vectors. In M. van Someren and G. Widmer, editors, Proc. ECML–97, pages 329–336. Springer-Verlag, 1997.

13. M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulates et axioms de l'école americaine. *Econometrica*, 21:503–546, 1953.

14. K. Althoff, S. Wess, and R. Traphoner. INRECA: A seamless integration of induction and case-based reasoning for design support. In *Proceedings 8th Workshop German SIG on Machine Learning*, 1995.

15. F.J. Anscombe and R.J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205, 1963.

16. K.J. Arrow and L. Hurwicz. An optimality criterion for decision making under ignorance. In C. Carter and J.L. Ford, editors, *Uncertainty and Expectations in Economics*. Basil Blackwell and Mott Ltd., Oxford, 1972.

17. C.G. Atkeson, A.W. Moore, and S. Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11:11–73, 1997.

18. E. Auriol, M. Manago, K.D. Althoff, S. Wess, and S. Dittrich. Integrating induction and case-based reasoning: Methodological approach and first evaluations. In J.P. Haton, M. Keane, and M. Manago, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-94, number 984 in LNAI, pages 18–32. Springer-Verlag, 1994.

19. R. Axelrod. *The Evolution of Cooperation*. Basic Books, Inc., New York, 1984.

20. T. Bailey and A.K. Jain. A note on distance-weighted k-nearest neighbor rules. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC–8(4):311–313, 1978.

21. H. Bandemer. Unscharfe Analyse unscharfer Daten. In R. Seising, editor, *Fuzzy Theorie und Stochastik*, pages 251–267. Vieweg, Wiesbaden, 1999.

22. H. Bandemer and W. Näther. *Fuzzy Data Analysis*. Kluwer Academic Publishers, Dordrecht, 1992.

23. E.B. Baum and W.D. Smith. A Bayesian approach to relevance in game playing. *Artificial Intelligence*, 97:195–242, 1997.

24. J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

25. D.A. Bell, J.W. Guan, and S.K. Lee. Generalized union and project operations for pooling uncertain and imprecise information. *Data & Knowledge Engineering*, 18:89–117, 1996.

26. R.E. Bellman, R. Kalaba, and L.A. Zadeh. Abstraction and pattern classification. *Journal of Mathematical Analysis and Applications*, 13:1–7, 1966.

27. M. Béreau and B. Dubuisson. A fuzzy extended k-nearest neighbors rule. *Fuzzy Sets and Systems*, 44:17–32, 1991.

28. J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2. edition, 1985.

29. R. Bergmann, K.D. Althoff, S. Breen, M. Göker, M. Manago, R. Traphöner, and S. Wess. *Developing industrial case-based reasoning applications: The INRECA methodology*, volume 1612 of *LNAI*. 2 edition, 2003.

30. R. Bergmann and W. Wilke. Towards a new formal model of transformational adaptation in case-based reasoning. In H. Prade, editor, ECAI-98*, 13th European Conference on Artificial Intelligence*, pages 53–57, 1998.

31. M. Berthold. Fuzzy logic. In M. Berthold and D.J. Hand, editors, *Intelligent Data Analysis*, pages 269–298. Springer-Verlag, Berlin, 1999.

32. J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithm.* Plenum Press, New York, 1981.

33. J.C. Bezdek, K. Chuah, and D. Leep. Generalized k-nearest neighbor rules. *Fuzzy Sets and Systems*, 18:237–256, 1986.

34. L. Bianco and G. Gerla. Logics with approximate premises. *International Journal of Intelligent Systems*, 13:1–10, 1998.

35. I. Bichindaritz, E. Kansu, and K.M. Sullivan. Case-based reasoning in Care-Partner: Gathering evidence for evidence-based medical practice. In B. Smyth and P. Cunningham, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-98*, 4th European Workshop on Case-Based Reasoning*, number 1488 in LNAI, pages 334–345. Springer-Verlag, 1998.

36. M. Blonski. Social learning with case-based decisions. *Journal of Economic Behaviour & Organization*, 38:59–77, 1999.

37. A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Occam's razor. *Information Processing Letters*, 24:377–380, 1987.

38. M. Boddy and T.L. Dean. Deliberation scheduling for problem solving in time-constrained environments. *Artificial Intelligence*, 67:245–285, 1994.

39. B. Bonet and H. Geffner. Arguing for decisions: a qualitative model for decision making. In *Proceedings* UAI-96, pages 98–105, 1996.

40. P. Bonissone and W. Cheetman. Financial applications of fuzzy case-based reasoning to residential property valuation. In *Proceedings of the 6th* IEEE *International Conference on Fuzzy Systems* (FUZZ-IEEE-97), pages 37–44, Barcelona, 1997.

41. C. Borgelt, J. Gebhardt, and R. Kruse. Possibilistic graphical models. In *Proceedings* ISSEK-98, Udine, Italy, 1998.

42. C. Borgelt and R. Kruse. Probabilistic and possibilistic networks and how to learn them from data. In O. Kaynak, L. Zadeh, B. Türksen, and I. Rudas, editors, *Soft Computing and Its Applications*, pages 403–426. Springer-Verlag, New York, 1998.

43. C. Borgelt and R. Kruse. *Graphical Models – Methods for Data Analysis and Mining.* Wiley, Chichester, 2002.

44. B. Bouchon-Meunier, editor. *Aggregation and Fusion of Imperfect Information.* Physica-Verlag, Heidelberg, 1998.

45. B. Bouchon-Meunier, J. Delechamp, C. Marsala, and M. Rifqi. Several forms of analogical reasoning. In *Proceedings* FUZZ-IEEE-97, pages 45–50, Barcelona, 1997.

46. B. Bouchon-Meunier, D. Dubois, L. Godo, and H. Prade. Fuzzy sets and possibility theory in approximate reasoning and plausible reasoning. In J.C. Bezdek, D. Dubois, and H. Prade, editors, *Fuzzy Sets in Approximate Reasoning and Information Systems*, pages 15–190. Kluwer, 1999.

47. B. Bouchon-Meunier, M. Rifqi, and S. Bothorel. Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84:143–153, 1996.

48. B. Bouchon-Meunier and L. Valverde. Analogy relations and inference. In *Proceedings 2nd* IEEE *International Conference on Fuzzy Systems*, pages 1140–1144, San Francisco, California, 1993.

49. B. Bouchon-Meunier and L. Valverde. A fuzzy approach to analogical reasoning. *Soft Computing*, 3:141–147, 1999.

50. C. Boutilier. Toward a logic for qualitative decision theory. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proceedings* KR-94*, 4th International Conference on Principles of Knowledge Representation and Reasoning*, pages 75–86, Bonn, Germany, 1994.

51. R. Bradley and N. Swartz. *Possible Worlds.* Basil Blackwell, Oxford, UK, 1979.

52. R. Brafmann and M. Tennenholtz. On the foundations of qualitative decision theory. In *Proceedings* AAAI-96*, 13th National Conference on Artificial Intelligence*, pages 1291–1296. AAAI-Press, 1996.

53. R. Brafmann and M. Tennenholtz. On the axiomatization of qualitative decision criteria. In *Proceedings* AAAI-97*, 14th National Conference on Artificial Intelligence*, pages 76–81. AAAI-Press, 1997.

54. J.S. Breese and D. Heckermann. Decision-theoretic case-based reasoning. In *Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics*, pages 56–63, Ft. Lauderdale, U.S.A., 1995.

55. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, 1984.

56. D.G. Bridge. Defining and combining symmetric and asymmetric similarity measures. In B. Smith and P. Cunningham, editors, *Advances in Case-Based Reasoning, Proceedings* EWCBR-98, number 1488 in LNAI, pages 52–63, Dublin, Ireland, 1998.

57. C.E. Brodley. Addressing the selective superiority problem: Automatic algorithm for model class selection. In *Proceedings 10th Machine Learning Conference*, pages 17–24, 1993.

58. H. Bunke and B.T. Messmer. Similarity measures for structured representations. In S. Wess, K.D. Althoff, and M.M. Richter, editors, *Topics in Case-Based Reasoning, Proceedings* EWCBR-94, number 837 in LNAI, pages 106–118. Springer-Verlag, 1994.

59. H.D. Burkhard. Extending some concepts of CBR - foundations of case retrieval nets. In M. Lenz, B. Bartsch-Spörl, H.D. Burkhard, and S. Wess, editors, *Case-Based Reasoning Technology*, number 1400 in Lecture Notes in Artificial Intelligence, pages 17–50. Springer-Verlag, 1998.

60. R. Carnap. A basic system of inductive logic, part 2. In R. Jeffrey, editor, *Studies in Inductive Logic and Probability, vol. II*, pages 7–155. University of California Press, Berkeley, 1980.

61. N. Cercone, A. An, and C. Chan. Rule-induction and case-based reasoning: hybrid architectures appear advantageous. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):166–174, 1999.

62. C.L. Chang. Finding prototypes for nearest neighbor classifiers. IEEE *Transactions on Computers*, C-23(11):1179–1184, 1974.

63. L. Chang and P. Harrison. A case-based reasoning testbed for experiments in adaptive memory retrieval and indexing. In D.W. Aha and A. Ram, editors, *Proceedings of the* AAAI *Fall Symposium on Adaptation of Knowledge for Reuse.* AAAI Press, 1995.

64. A.R. Chaturvedi and G.K. Hutchinson AndD.L. Nazareth. Supporting complex real-time decision making through machine learning. *Decision Support Systems*, 10:213–233, 1993.

65. W. Cheetham. Case-based reasoning with confidence. In *EWCBR–2000, 5th European Workshop on Case-Based Reasoning*, pages 15–25, Trento, Italy, 2000. Springer-Verlag.

66. W. Cheetham and J. Price. Measures of solution accuracy in case-based reasoning systems. In *Proc. ECCBR–2004, 7th European Conference on Case-Based Reasoning*, pages 106–118, Madrid, Spain, 2004. Springer-Verlag.

67. G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.

68. C.K. Chow. On optimum recognition error and reject tradeoff. IEEE *Transactions on Information Theory*, IT-16:41–46, 1970.

69. M. Cohen and J.Y. Jaffray. Rational behavior under complete ignorance. *Econometrica*, 48:1281–1299, 1980.

70. M. Cohen and J.Y. Jaffray. Decision making in a case of mixed uncertainty: A normative model. *Journal of Mathematical Psychology*, 29(4):428–434, 1985.

71. R. Comolli. Is case-based decision theory rational? *Rivista Internazionale Di Scienze Economiche E Commerciali*, 45(1):99 –114, 1998.

72. R.M. Cooke. *Experts in Uncertainty.* Oxford University Press, London, 1991.

73. T.M. Cover. Estimation by the nearest neighbor rule. IEEE *Transactions of Information Theory*, IT-14(1):50–55, 1968.

74. T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. IEEE *Transactions on Information Theory*, IT-13:21–27, 1967.

75. B.V. Dasarathy. Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(1):67–71, 1980.

76. B.V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, California, 1991.

77. B.V. Dasarathy. NN concepts and techniques. An introductory survey. In B.V. Dasarathy, editor, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, pages 1–30. IEEE Computer Society Press, Washington, 1991.

78. E.R. Davies. Training sets and a priori probabilities with the nearest neighbor method of pattern classification. *Pattern Recognition Letters*, 8(1):11–13, 1988.

79. R. Lopez de Mantaras and E. Armengol. Machine learning from examples: Inductive and lazy methods. *Data & Knowledge Engineering*, 25:99–123, 1998.

80. R. Lopez de Mantaras and E. Plaza. Case-based reasoning: An overview. *AI Communications Journal*, 10(1):21–29, 1997.

81. SJ. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh. Generating estimates of classification confidence for a case-based spam filter. In *Proc. ICCBR-2005, 6th International Conference on Case-Based Reasoning*, pages 177–190, Chicago, Illinois, 2005. Springer-Verlag.

82. A.P. Dempster. Upper and lower probability induced by a random closed interval. *Annals of Mathematical Statistics*, 39:219–246, 1968.

83. D. Denneberg. *Non-Additive Measure and Integral*. Kluwer Academic Publishers, 1994.

84. T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer Theory. IEEE *Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.

85. T. Denoeux and L.M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):409–424, 2001.

86. P. Diamond and P. Kloeden. *Metric Spaces of Fuzzy Sets: Theory and Applications*. World Scientific, Singapur, 1994.

87. P. Diamond and H. Tanaka. Fuzzy regression analysis. In R. Slowinski, editor, *Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, pages 349–387. Kluwer, 1998.

88. J.K. Dixon. Pattern recognition with partly missing data. IEEE *Transactions on Systems, Man, and Cybernetics*, 9(10):617–621, 1979.

89. P. Domingos. Rule induction and instance-based learning: A unified approach. In C.S. Mellish, editor, *Proceedings* IJCAI-95*, 14th International Joint Conference on Artificial Intelligence*, pages 1226–1232, Montreal, 1995. Morgan Kaufmann.

90. P. Domingos. Unifying instance-based and rule-based induction. *Machine Learning*, 24:141–168, 1996.

91. P. Domingos. The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3:409–425, 1999.

92. P. Domingos and G. Hulten. A general framework for mining massive data streams. *Journal of Computational and Graphical Statistics*, 12:945–949, 2003.

93. J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In A. Prieditis and S. Russell, editors, *Machine Learning: Proceedings of the 12th International Conference*, pages 194–202. Morgan Kaufmann, 1995.

94. J. Doyle. What is rational psychology? *AI Magazine*, 4(3):50–53, 1983.

95. J. Doyle and T. Dean. Strategic directions in artificial intelligence. *AI Magazine*, 18(1):87–101, 1997.

96. Jon Doyle. Prospects for preferences. *Computational Intelligence*, 20(2):111–136, 2004.

97. D. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, 57:45–97, 1995.

98. D. Dubois, D. de Berre, H. Prade, and R. Sabbadin. Using possibilistic logic for modelling qualitative decision: ATMS-based algorithms. *Fundamenta Informaticae*, 37:1–30, 1999.

99. D. Dubois, F. Esteva, P. Garcia, L. Godo, R. Lopez de Mantaras, and H. Prade. Fuzzy set modelling in case-based reasoning. *International Journal of Intelligent Systems*, 13:345–373, 1998.

100. D. Dubois, F. Esteva, P. Garcia, L. Godo, R. Lopez de Mantaras, and H. Prade. Case-based reasoning: a fuzzy approach. In A.L. Ralescu and J.G. Shanahan, editors, *Proceedings* IJCAI-97 *Workshop on Fuzzy Logic in Artificial Intelligence*, number 1566 in Lecture Notes in Artificial Intelligence, pages 79–90. Springer-Verlag, 1999.

101. D. Dubois, F. Esteva, P. Garcia, L. Godo, R.L. de Mantaras, and H. Prade. Fuzzy modelling of case-based reasoning and decision. In D.B. Leake and E. Plaza, editors, *Case-based Reasoning Research and Development, Proceedings* ICCBR-97, pages 599–610. Springer-Verlag, 1997.

102. D. Dubois, F. Esteva, P. Garcia, L. Godo, and H. Prade. A logical approach to interpolation based on similarity relations. *International Journal of Approximate Reasoning*, 17(1):1–36, 1997.

103. D. Dubois, L. Godo, H. Prade, and A. Zapico. Possibilistic representation of qualitative utility: an improved characterization. In *Proceedings* IPMU-98*, 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 180–187, Paris, La Sorbonne, 1998. Editions E.D.K.

104. D. Dubois, P. Hajek, and H. Prade. Knowledge driven vs. data driven logics. *Journal of Logic, Language and Information*, 9:65–89, 2000.

105. D. Dubois, E. Hüllermeier, and H. Prade. Fuzzy set-based methods in instance-based reasoning. IEEE *Transactions on Fuzzy Systems*, 10(3):322–332, 2002.

106. D. Dubois, E. Hüllermeier, and H. Prade. On the representation of fuzzy rules in terms of crisp rules. *Information Sciences*, 151:301–326, 2003.

107. D. Dubois, E. Hüllermeier, and H. Prade. Formalizing case-based inference using fuzzy rules. In S.K. Pal, D.Y. So, and T. Dillon, editors, *Soft Computing in Case-Based Reasoning*, pages 47–72. Springer-Verlag, 2000.

108. D. Dubois, S. Kaci, and H. Prade. Bipolarity in reasoning and decision: An introduction. The case of the possibility framework. In *IPMU–04, 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Perugia, Italy, 2004.

109. D. Dubois, M. Nakata, and H. Prade. Extended divisions for flexible queries in relational databases. Technical Report 97-43 R, IRIT – Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, September 1997.

110. D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In M.M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-Holland, Amsterdam, 1982.

111. D. Dubois and H. Prade. Fuzzy sets and statistical data. *European Journal of Operational Research*, 25:345–356, 1986.

112. D. Dubois and H. Prade. Weighted minimum and maximum operations in fuzzy set theory. *Information Sciences*, 39:205–210, 1986.

113. D. Dubois and H. Prade. The principle of minimum specificity as a basis for evidential reasoning. In B. Bouchon and R.R. Yager, editors, *Uncertainty in Knowledge-Based Systems*, number 286 in Lecture Notes in Computer Science, pages 75–84. Springer-Verlag, Berlin, 1987.

114. D. Dubois and H. Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy Sets and Systems*, 24:161–182, 1987.

115. D. Dubois and H. Prade. On the combination of uncertain or imprecise pieces of information in rule-based systems - A discussion in the framework of possibility theory. *International Journal of Approximate Reasoning*, 2(1):65–87, 1988.

116. D. Dubois and H. Prade. *Possibility Theory.* Plenum Press, 1988.

117. D. Dubois and H. Prade. A typology of fuzzy "if ... then ..." rules. In *Proceedings of the 3rd International Fuzzy Systems Association (IFSA) Congress*, pages 782–785, Seattle, WA, 1989.

118. D. Dubois and H. Prade. Fuzzy sets in approximate reasoning, part 1: Inference with possiblity distributions. *Fuzzy Sets and Systems*, 40:143–202, 1991.

119. D. Dubois and H. Prade. Gradual inference rules in approximate reasoning. *Information Sciences*, 61(1,2):103–122, 1992.

120. D. Dubois and H. Prade. On the combination of evidence in various mathematical frameworks. In J. Flamm and T. Luisi, editors, *Reliability Data Collection and Analysis*, pages 213–241. Kluwer Academic Publishers, 1992.

121. D. Dubois and H. Prade. Possibility theory as a basis for preference propagation in automated reasoning. In FUZZ-IEEE-92*, Proceedings 1st* IEEE *Int. Conference on Fuzzy Systems*, pages 821–832, San Diego, Ca., 1992.

122. D. Dubois and H. Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

123. D. Dubois and H. Prade. Possibility theory as a basis for qualitative decision theory. In *Proceedings* IJCAI-95*, 14th International Joint Conference on Artificial Intelligence*, pages 1924–1930, Montreal, 1995.

124. D. Dubois and H. Prade. What are fuzzy rules and how to use them. *Fuzzy Sets and Systems*, 84:169–185, 1996.

125. D. Dubois and H. Prade. A fuzzy set approach to case-based decision. In R. Felix, editor, EFDAN-97*, 2nd European Workshop on Fuzzy Decision Analysis and Neural Networks for Management, Planning and Optimization*, pages 1–9, Dortmund, Germany, 1997.

126. D. Dubois and H. Prade. The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90(2):141–150, 1997.

127. D. Dubois and H. Prade. Possibility theory: Qualitative and quantitative aspects. In D.M. Gabbay and P. Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 1*, pages 169–226. Kluwer Academic Publishers, 1998.

128. D. Dubois, H. Prade, and R. Sabbadin. Decision-theoretic foundations of qualitative possibility theory. *European Journal of Operational Research*, 128:459–478, 2001.

129. D. Dubois, H. Prade, and R. Sabbadin. Qualitative decision theory with Sugeno integrals. In *Proceedings* UAI-94*, 14th Conference on Uncertainty in Artificial Intelligence*, pages 121–128. Morgan Kaufmann, 1998.

130. D. Dubois, H. Prade, and P. Smets. Representing partial ignorance. IEEE *Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, 26(3):361–377, 1996.

131. D. Dubois, H. Prade, and P. Smets. Not impossible vs. guaranteed possible in fusion and revision. In *Proceedings* ESCQARU–2001, number 2143 in LNCS, pages 522–531, Toulouse, France, 2001. Springer-Verlag.

132. D. Dubois, H. Prade, and C. Testemale. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28:313–331, 1988.

133. D. Dubois, H. Prade, and L. Ughetto. Checking the coherence and redundancy of fuzzy knowledge bases. *IEEE Transactions on Fuzzy Systems*, 5(3):398–417, 1997.

134. D. Dubois, H. Prade, and L. Ughetto. A new perspective on reasoning with fuzzy rules. In N.R. Pal and M. Sugeno, editors, *Advances in Soft Computing, Proc. of the AFSS International Conference on Fuzzy Systems*, number 2275 in LNAI, pages 1–11, Calcutta, India, 2002. Springer-Verlag.

135. D. Dubois, H. Prade, and R.R. Yager. Merging fuzzy information. In J.C. Bezdek, D. Dubois, and H. Prade, editors, *Fuzzy Sets in Approximate Reasoning and Information Systems*, pages 335–401. Kluwer Academic Publishers, Boston, 1999.

136. B. Dubuisson and M. Masson. A statistical decision rule with incomplete knoweldge about classes. *Pattern Recognition*, 26(1):155–165, 1993.

137. S.A. Dudani. The distance-weighted k-nearest-neighbor rule. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327, 1976.

138. S. Dutta and P. Bonissone. Integrating case- and rule-based reasoning. *International Journal of Approximate Reasoning*, 8:163–203, 1993.

139. A.W.F. Edwards. *Likelihood*. Cambridge University Press, Cambridge, UK, 1972.

140. D. Ellsberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

141. F. Esteva, P. Garcia, L. Godo, and R. Rodriguez. A modal account of similarity-based reasoning. *International Journal of Approximate Reasoning*, 16:235–260, 1997.

142. B. Faltings. Probabilistic indexing for case-based prediction. In D.B. Leake and E. Plaza, editors, *Case-based Reasoning Research and Developement, Proceedings* ICCBR-97, pages 611–622. Springer-Verlag, 1997.

143. H. Fargier, J. Lang, and T. Schiex. Mixed constraint satisfaction: a framework for decision problems under incomplete knowledge. In *Proceedings* AAAI-96*, 13th National Conference on Artificial Intelligence*, pages 175–180, Portland, Oregon, 1996.

144. H. Farreny and H. Prade. About flexible matching and its use in analogical reasoning. In ECAI-82*, European Conference on Artificial Intelligence*, pages 43–47, Orsay, France, July 1982.

145. T.L. Fine. *Theories of Probability.* Adademic Press, New York, 1973.

146. B. De Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Insitut Henri Poincaré*, VII:1–68, 1937.

147. R.A. Fisher. *Statistical Methods and Bayesian Inference.* Oliver and Boyd, Edinburgh, 1956.

148. E. Fix and J.L. Hodges. Discriminatory analysis: nonparametric discrimination: consistency principles. In B.V. Dasarathy, editor, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, California, 1991. Reprint of original work from 1951.

149. E. Fix and J.L. Hodges. Discriminatory analysis: nonparametric discrimination: small sample performance. In B.V. Dasarathy, editor, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* IEEE Computer Society Press, Los Alamitos, California, 1991. Reprint of original work from 1952.

150. J. Fodor. Contrapositive symmetry of fuzzy implications. *Fuzzy Sets and Systems*, 69(2):141–156, 1995.

151. J. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

152. P. Fortemps and M. Pirlot. Axiomatic characterization of some ordinal choice procedures. In B. De Baets, J. Fodor, and L.T. Kóczy, editors, *Proceedings* EUROFUSE-SIC-99, pages 122–125, Budapest, Hungary, 1999.

153. S. French. Group consensus probability distributions: A critical survey. In J.M. Bernardo et. al., editor, *Bayesian Statistics 2*, pages 183–201. North Holland, Amsterdam, 1985.

154. J.H. Friedman, F. Baskett, and L.J. Shustek. An algorithm for finding nearest neighbors. IEEE *Transactions on Computers*, 24:1000–1006, 1975.

155. J.H. Friedman, R. Kohavi, and Y. Yun. Lazy decision trees. In *Proceedings* AAAI–96, pages 717–724, Menlo Park, California, 1996. Morgan Kaufmann.

156. J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In *Proceedings* ECML–2003*, 13th European Conference on Machine Learning*, pages 145–156, Cavtat-Dubrovnik, Croatia, September 2003. Springer-Verlag.

157. K. Fukunaga and T. Flick. A parametrically-defined nearest neighbor distance measure. *Pattern Recognition Letters*, 1:3–5, 1982.

158. K. Fukunaga and P.M. Narendra. A branch and bound algorithm for computing k-nearest neighbors. IEEE *Transactions on Computers*, 24:750–753, 1975.

159. T. Gabel and M. Riedmiller. CBR for state value function approximation in reinforcement learning. In *Proc. ICCBR–2005, 6th International Conference on Case-Based Reasoning*, pages 206–221, Chicago, Illinois, 2005.

160. T. Gabel and A. Stahl. Exploiting background knowledge when learning similarity measures. In *ECCBR-2004, 7th European Conference on Case-Based Reasoning*, pages 169–183, Madrid, Spain, 2004.

161. MM. Gaber and A. Zaslavsky amd S. Krishnaswamy. Mining data streams: A review. *ACM SIGMOD Record*, 34(1), 2005.

162. A. Gammerman and V. Vovk. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science*, 287:209–217, 2002.

163. G.W. Gates. The reduced nearest neighbor rule. IEEE *Transactions on Information Theory*, IT-18:431–433, 1972.

164. J. Gebhardt and R. Kruse. Parallel combination of information sources. In D.M. Gabbay and Ph. Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 3*, pages 393–439. Kluwer Academic Publishers, 1998.

165. C. Genest and J.V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–148, 1986.

166. I. Gilboa. Expected utility with purely subjective non-additive probability. *Journal of Mathematical Economics*, 16:65–88, 1987.

167. I. Gilboa and D. Schmeidler. Case-based decision theory. *Quarterly Journal of Economics*, 110(4):605–639, 1995.

168. I. Gilboa and D. Schmeidler. Case-based optimization. *Games and Economic Behavior*, 15(1):1–26, 1996.

169. I. Gilboa and D. Schmeidler. Act similarity in case-based decision theory. *Economic Theory*, 9:47–61, 1997.

170. I. Gilboa and D. Schmeidler. Cumulative utility consumer theory. *International Economic Review*, 38:737–761, 1997.

171. I. Gilboa and D. Schmeidler. Case-based decision: An extended abstract. In H. Prade, editor, *Proceedings* ECAI-98*, 13th European Conference on Artificial Intelligence*, pages 706–710, Brighton, UK, 1998.

172. I. Gilboa and D. Schmeidler. Case-based knowledge and induction. IEEE *Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, 30(2):85–95, 2000.

173. I. Gilboa and D. Schmeidler. Inductive inference: An axiomatic approach. *Econometrica*, 71:1–26, 2003.

174. A.R. Golding and P.S. Rosenbloom. Improving rule-based systems through case-based reasoning. In *Proceedings* AAAI-91, pages 22–27, 1991.

175. A.R. Golding and P.S. Rosenbloom. Improving accuracy by combining rule-based and case-based reasoning. *Artificial Intelligence*, 87:215–254, 1996.

176. A.J. Gonzales and R. Laureano-Oritz. A case-based reasoning approach to real estate property appraisal. *Expert Systems with Applications*, 4:229–246, 1992.

177. I.R. Goodman. Fuzzy sets as equivalence classes of random sets. In R.R. Yager, editor, *Fuzzy Sets and Possibility Theory*, pages 327–342. Pergamon Press, Oxford, 1982.

178. N. Goodman. Seven strictures on similarity. In N. Goodman, editor, *Problems and Projects.* Bobbs-Merrill, New York, 1972.

179. M. Grabisch. Fuzzy integral for classification and feature extraction. In M. Grabisch, T. Murofushi, and M. Sugeno, editors, *Fuzzy Measures and Integrals: Theory and Applications*, pages 415–434. Physica-Verlag, 2000.

180. M. Grabisch, H.T. Nguyen, and E.A. Walker. *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference.* Kluwer Academic Publishers, 1995.

181. Vu Ha and P. Haddawy. Similarity of personal preferences: theoretical foundations and empirical analysis. *Artificial Intelligence*, 146:149–173, 2003.

182. I. Hacking. Slightly more realistic personal probabilities. *Philosophical Science*, 34:311–325, 1967.

183. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* MIT Press, Cambridge, 2001.

184. S. Har-Peled, D. Roth, and D. Zimak. Constraint classification: a new approach to multiclass classification. In *Proceedings 13th Int. Conf. on Algorithmic Learning Theory*, pages 365–379, Lübeck, Germany, 2002. Springer.

185. W. Härdle and M. Müller. Nichtparametrische Glättungsmethoden in der alltäglichen Praxis. *Allgemeines Statistisches Archiv*, 77:9–31, 1993.

186. P.E. Hart. The condensed nearest neighbor rule. IEEE *Transactions on Information Theory*, IT-14:515–516, 1968.

187. D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

188. M.E. Hellman. The nearest neighbor classification rule with a reject option. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-6:179–185, 1970.

189. K. Hirota. Concepts of probabilistic sets. *Fuzzy Sets and Systems*, 5:31–46, 1981.

190. C.S. Hong and P. Wakker. The comonotonic sure-thing principle. *Journal of Risk and Uncertainty*, 12:5–27, 1996.

191. F. Höppner, F. Klawonn, F. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. Wiley, Chichester, 1999.

192. E.J. Horvitz. Reasoning about beliefs and actions under computational resource constraints. In L.N. Kanal, T.S. Levitt, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 3*, pages 301–324. North-Holland, Amsterdam, 1989.

193. E.J. Horvitz, J.S. Breese, and M. Henrion. Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2:247–302, 1988.

194. J.L. Hougaard and H. Keiding. Representation of preferences on fuzzy measures by a fuzzy integral. *Mathematical Social Science*, 31:1–17, 1996.

195. Y. Huang. An evolutionary agent model of case-based classification. In B. Smith and B. Faltings, editors, *Advances in Case-Based Reasoning, Proceedings* EWCBR-96*, 3rd European Workshop on Case-Based Reasoning*, number 1168 in LNAI, pages 193–203, Lausanne, 1996. Springer-Verlag.

196. E. Hüllermeier. Experience-based decision making: A satisficing decision tree approach. IEEE *Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 35(5):641–653, 2005.

197. E. Hüllermeier. Possibilistic instance-based learning. *Artificial Intelligence*, 148(1–2):335–383, 2003.

198. E. Hüllermeier. Cho-k-NN: A method for combining interacting pieces of evidence in case-based learning. In L. Kaelbling and A. Saffiotti, editors, *Proceedings* IJCAI–05*, 19th International Joint Conference on Artificial Intelligence*, pages 3–8, Edinburgh, Scotland, 2005.

199. E. Hüllermeier. Toward a probabilistic formalization of case-based inference. In T. Dean, editor, *Proceedings* IJCAI–99*, 16th International Joint Conference on Artificial Intelligence*, pages 248–253, Stockholm, Sweden, July/August 1999. Morgan Kaufmann.

200. E. Hüllermeier. Focusing search by using problem solving experience. In W. Horn, editor, *Proceedings* ECAI–2000*, 14th European Conference on Artificial Intelligence*, pages 55–59, Berlin, Germany, 2000. IOS Press.

201. E. Hüllermeier. Instance-based prediction with guaranteed confidence. In R. Lopez de Mantaras and L. Saitta, editors, *Proceedings* ECAI–2004*, 16th European Conference on Artificial Intelligence*, pages 97–101, Valencia, Spain, 2004. IOS Press.

202. E. Hüllermeier. Similarity-based inference as evidential reasoning. In W. Horn, editor, *Proceedings* ECAI–2000*, 14th European Conference on Artificial Intelligence*, pages 50–54, Berlin, Germany, 2000. IOS Press.

203. E. Hüllermeier. On the representation and combination of evidence in instance-based learning. In *Proceedings* ECAI–2002*, 15th European Conference on Artificial Intelligence*, pages 360–364, Lyon, France, 2002. IOS Press.

204. E. Hüllermeier and J. Beringer. Learning decision rules from positive and negative preferences. In *IPMU–04, 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 967–974, Perugia, Italy, 2004.

205. E. Hüllermeier, D. Dubois, and H. Prade. Fuzzy rules in case-based reasoning. In *Conférences* AFIA–99*, Proceedings* RÀPC–99*, Raisonnement à partir de Cas*, pages 45–54, Paris, Palaiseau, June 1999.

206. D. Hume. *An Enquiry concerning Human Understanding*. Oxford University Press Inc., New York, 1999.

207. R. Hummel and L. Landy. Evidence as opinions of experts. In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 43–53. North-Holland, 1988.

208. K. Jabbour, J. Riveros, D. Landsbergen, and W. Meyer. ALFA: Automated load forecasting assistant. IEEE *Transactions on Power Apparatus and Systems*, 3(3):908–914, 1988.

209. M. Jaczynski and B. Trousse. Fuzzy logic for the retrieval step of a case-based reasoner. In EWCBR-94*, Proceedings of the European Workshop on Case-Based Reasoning*, pages 313–321, 1994.

210. J.Y. Jaffray. Linear utility theory for belief functions. *Operations Research Letters*, 8:107–112, 1989.

211. J.Y. Jaffray. Dynamic decision making with belief functions. In R.R. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer Theory of Evidence*, pages 331–352. Wiley, 1992.

212. J.Y. Jaffray and P. Wakker. Decision making with belief functions: Compatibility and incompatibility with the sure-thing principle. *Journal of Risk and Uncertainty*, 8:255–271, 1994.

213. K.P. Jantke. Nonstandard concepts of similarity in case-based reasoning. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis: Prospects, Foundations, Applications*, pages 29–44. Springer-Verlag, 1994.

214. B.C. Jeng and T.P. Liang. Fuzzy indexing and retrieval in case-based systems. *Expert Systems with Applications*, 8(1):135–142, 1995.

215. X. Jiang, A. MuEnger, and H. Bunke. On median graphs: Properties, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 2001.

216. A. Józwik. A learning scheme for a fuzzy k-NN rule. *Pattern Recognition Letters*, 1:287–289, 1983.

217. S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, and D.W. Aha. A probabilistic framework for memory-based reasoning. *Artificial Intelligence*, 104(1-2):287–311, 1998.

218. J.M. Keller, M.R. Gray, and J.A. Givens. A fuzzy k-nearest neighbor algorithm. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-15(4):580–585, 1985.

219. F. Kerestecioglu. *Change Detection and Input Design in Dynamical Systems*. John Wiley & Sons Inc., 1993.

220. D. Kibler and D.W. Aha. Learning representative exemplars of concepts: An initial study. In *Proceedings of the Fourth International Workshop on Machine Learning*, pages 24–29, UC-Irvine, 1987.

221. D. Kibler, D.W. Aha, and M.K. Albert. Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5:51–57, 1989.

222. B.S. Kim and S.B. Park. A fast $k$ nearest neighbor finding algorithm based on the ordered partition. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 8(6):761–766, 1985.

223. J. Kittler. A method for determining k-nearest neighbors. *Kybernetes*, 7:313–315, 1978.

224. J. Kittler. Feature selection and extraction. In T.Y. Young and K.S. Fu, editors, *Handbook of Pattern Recognition and Image Processing*, pages 59–81. Academic Press, 1986.

225. F. Klawonn and J.L. Castro. Similarity in fuzzy reasoning. *Mathware & Soft Computing*, 2:197–228, 1995.

226. F. Klawonn, J. Gebhardt, and R. Kruse. Fuzzy control on the basis of equality relations with an example from idle speed control. *IEEE Transactions on Fuzzy Systems*, 3(3):336–350, 1995.

227. EP. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer Academic Publishers, 2002.

228. G.J. Klir. *Facets of Systems Science*. Plenum Press, 1991.

229. G.J. Klir. Measures of uncertainty in the Dempster-Shafer theory of evidence. In R.R. Yager, M. Fedrizzi, and J. Kacprzyk, editors, *Advances in the Dempster-Shafer theory of evidence*, pages 35–49. Wiley, New York, 1994.

230. G.J. Klir and T.A. Folger. *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, 1988.

231. G.J. Klir and M.J. Wierman. *Uncertainty-Based Information*. Physica-Verlag, Heidelberg, 1998.

232. G.J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic. Theory and Applications*. Prentice-Hall, New York, 1995.

233. J.L. Kolodner, editor. DAPRA-88*, Workshop on case-based reasoning*. Morgan Kaufmann, San Mateo, 1988.

234. J.L. Kolodner. *Case-based Reasoning*. Morgan Kaufmann, San Mateo, 1993.

235. P. Kontkanen, J. Lahtinen, P. Myllymäki, and H. Tirri. An unsupervised Bayesian distance measure. In E. Blanzieri and L. Portinale, editors, *Advances in Case-Based Reasoning, Proceedings* EWCBR-2000, pages 148–160. Springer-Verlag, 2000.

236. P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Bayes optimal instance-based learning. In B. Smyth and P. Cunningham, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-98*, 4th European Workshop on Case-Based Reasoning*, number 1488 in LNAI, pages 77–88. Springer-Verlag, 1998.

237. P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Bayes optimal instance-based learning. In B. Smyth and P. Cunningham, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-98*, 4th European Workshop on Case-Based Reasoning*, number 1488 in LNAI, pages 77–88. Springer-Verlag, 1998.

238. D.R. Kraay and P.T. Harker. Case-based reasoning for repetitive combinatorial optimization problems, part I: Framework. *Journal of Heuristics*, 2:55–85, 1996.

239. D.R. Kraay and P.T. Harker. Case-based reasoning for repetitive combinatorial optimization problems, part II: Numerical results. *Journal of Heuristics*, 3:25–42, 1997.

240. R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. IEEE *Transactions on Fuzzy Systems*, 1(2):98–110, 1993.

241. R. Kruse and D. Meyer. *Statistics with Vague Data*. D. Reidel, Dordrecht, 1987.

242. B.J. Kuipers. *Qualitative Reasoning*. MIT Press, 1994.

243. M. Kukar and I. Kononenko. Reliable classifications with machine learning. In *Proc. European Conference on Machine Learning, ECML*, pages 219–231, 2002.

244. G. Lakoff. Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2:458–508, 1973.

245. S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

246. D.B. Leake. CBR in context: the present and the future. In D.B. Leake, editor, *Case-Based Reasoning: Experiences, Lessons, and Future Directions*, pages 3–30. AAAI Press, Menlo Park, California, 1996.

247. DB. Leake and DC. Wilson. When experience is wrong: Examining CBR for changing tasks and environments. In *Proc. ICCBR–99*, pages 218–232, 1999.

248. M.H. Lee. On models, modelling and the distinctive nature of model-based reasoning. *AI Communications*, 12:127–137, 1999.

249. M. Lenz, B. Bartsch-Spörl, H.D. Burkhard, and S. Wess, editors. *Case-Based Reasoning Technology.* Springer-Verlag, 1998.

250. D. Lewis. *Counterfactuals.* Basil Blackwell, 1993.

251. D.K. Lewis. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2:418–446, 1973.

252. T.W. Liao and Z. Zhang. Similarity measures for retrieval in case-based reasoning. *Applied Artificial Intelligence*, 12:267–288, 1998.

253. J. Lieber. A criterion of comparison between two case-bases. In *Proceedings* EWCBR-95*, 2nd European Workshop on Case-Based Reasoning*, pages 87–100, 1995.

254. M. Lindenbaum, S. Marcovich, and D. Rusakov. Selective sampling for nearest neighbor classifiers. In *Proceedings* AAAI-99*, 16th National Conference on Artificial Intelligence*, pages 366–371, Orlando, Florida, 1999.

255. B.L. Lipman. How to decide how to decide how to ...: Modeling limited rationality. *Econometrica*, 59(4):1105–1125, 1991.

256. D.O. Loftsgaarden and C.P. Quesenberry. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, 36:1049–1051, 1965.

257. J. Macleod, A. Lik, and D. Titterington. A re-examination of the distance-weighted k-nearest neighbor classification rule. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC–17(4):689–696, 1987.

258. E. Mamdani. Application of fuzzy logic to approximate reasoning using linguistic systems. IEEE *Transactions on Computers*, 26:1182–1191, 1977.

259. E. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7:1–13, 1975.

260. J.G. March and H.A. Simon. *Organizations.* Wiley, New York, 1958.

261. O. Maron. Using errors to create piecewise learnable partitions. In *Proceedings* AAAI-94, pages 1474–1479, 1994.

262. A. Matsui. Expected utility theory and case-based reasoning. *Mathematical Social Sciences*, 39:1–12, 2000.

263. S.H. McIntyre, D.D. Achabal, and C.M. Miller. Applying case-based reasoning to forecasting retail sales. *Journal of Retailing*, 69(4):372–398, 1993.

264. E. McKenna and B. Smyth. Competence-guided edition methods for lazy learning. In *Proceedings* ECAI–2000*, 14th European Conference on Artificial Intelligence*, pages 60–64, Berlin, 2000.

265. D. McSherry. An adaptation heuristic for case-based estimation. In B. Smyth and P. Cunningham, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-98, number 1488 in LNAI, pages 184–195. Springer-Verlag, 1998.

266. J.S. Mill. *A System of Logic.* Longmans, Green, and Co., 1906. (Original edition 1843).

267. J. Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4:227–243, 1989.

268. J. Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3:319–342, 1989.

269. T.M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings* IJCAI-77, pages 305–310, 1977.

270. T.M. Mitchell. The need for biases in learning generalizations. Technical Report TR CBM–TR–117, Rutgers University, 1980.

271. T.M. Mitchell. *Machine Learning*. McGraw-Hill, Boston, Massachusetts, 1997.

272. S. Moral and J. del Sagrado. Aggregation of imprecise probabilities. In B. Bouchon-Meunier, editor, *Aggregation and Fusion of Imperfect Information*, pages 162–168. Physica-Verlag, Heidelberg, 1998.

273. A. Mosleh and G. Apostolakis. Models for the use of expert opinions. In R.A. Waller and V.T. Covello, editors, *Low Probability/High Consequence Risk Analysis*. Plenum Press, New York, 1984.

274. P. Myllymäki and H. Tirri. Bayesian case-based reasoning with neural networks. In *Proceedings* IEEE *International Conference on Neural Networks*, pages 422–427, San Francisco, 1993.

275. P. Myllymäki and H. Tirri. Massively parallel case-based reasoning with probabilistic similarity metrics. In K.D. Althoff, S. Weiss, and M.M. Richter, editors, *Topics in Case-Based Reasoning*, number 837 in Lecture Notes in Artificial Intelligence, pages 144–154. Springer-Verlag, 1994.

276. G. Nakhaeizadeh. Learning the prediction of time series. A theoretical and empirical comparison of CBR with some other approaches. In S. Wess, K.D. Althoff, and M.M. Richter, editors, *Proceedings* EWCBR-93*, First European Workshop on Case-Based Reasoning*, pages 66–76. Springer-Verlag, 1993.

277. D. Nauck and R. Kruse. Neuro-fuzzy methods in fuzzy rule generation. In J.C. Bezdek, D. Dubois, and H. Prade, editors, *Fuzzy Sets in Approximate Reasoning and Information Systems*, The Handbooks of Fuzzy Sets Series, pages 305–334. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

278. J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* John Wiley and Sons, 1953.

279. J. Neyman. On a new class of 'contagious' distributions. *Annals of Mathematical Statistics*, 10:35–57, 1939.

280. H.T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

281. I. Niiniluoto. Analogy and similarity in scientific reasoning. In D.H. Helman, editor, *Analogical Reasoning*, pages 271–298. Kluwer Academic Publishers, Dordrecht, 1988.

282. D. O'Leary. Verification and validation of case-based systems. *Expert Systems with Applications*, 6:57–66, 1993.

283. H. Osborne and D. Bridge. Similarity metrics: A formal unification of cardinal and non-cardinal similarity measures. In *Proceedings* ICCBR-97*, 2nd International Conference on Case-Based Reasoning*, pages 235–244, 1997.

284. S.V. Ovchinnikov. Similarity relations, fuzzy partitions, and fuzzy orderings. *Fuzzy Sets and Systems*, 40:107–126, 1991.

285. SK. Pal, TS. Dillon, and DS. Yeung, editors. *Soft Computing in Case Based Reasoning*. Springer-Verlag, 2001.

286. R. Pan, Q. Yang, JJ. Pan, and L. Li Competence. Driven Case-Base Mining. In *Proc. AAAI–2005*, 228-233. Pittsburgh, USA.

287. AN. Papadopoulos and Y. Manolopoulos. *Nearest Neighbor Search: A Database Perspective*. Series in Computer Science. Springer-Verlag, 2005.

288. G. Parthasarathy and B.N. Chatterji. A class of new KNN methods for low sample problems. IEEE *Transactions on Systems, Man, and Cybernetics*, 20(3):715–718, 1990.

289. E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

290. E.A. Patrick and F.P. Fischer. A generalized k-nearest neighbor rule. *Information and Control*, 16(2):128–152, 1970.

291. J. Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, 1984.

292. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

293. J. Pearl. From qualitative utility to conditional "ought to". In D. Heckerman and H. Mamdani, editors, *Proceedings 9th International Conference on Uncertainty in Artificial Intelligence*, pages 12–20, San Mateo, CA, 1993. Morgan Kaufmann.

294. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

295. E. Plaza and R. Lopez de Mantaras. A case-based apprentice that learns from fuzzy examples. In Z. Ras, M. Zemankova, and M. Emrich, editors, *Methodologies for Intelligent Systems*, pages 420–427. Elsevier, 1990.

296. E. Plaza, F. Esteva, P. Garcia, L. Godo, and R. Lopez de Mantaras. A logical approach to case-based reasoning using fuzzy similarity relations. *Information Sciences*, 106:105–122, 1998.

297. G. Pólya. *Mathematics and Plausible Reasoning, Volume 2: Patterns of Plausible Inference*. Princeton University Press, Princeton, New Jersey, 1954.

298. J.C. Pomerol. Artificial intelligence and human decision making. *European Journal of Operational Research*, 99:3–25, 1997.

299. K.R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1974.

300. H. Prade and R.R. Yager. Estimations of expectedness and potential surprize in possibility theory. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:417–428, 1994.

301. K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition. In *Proc. European Conference on Machine Learning, ECML*, pages 381–390, 2002.

302. F. Puppe. *Systematic Introduction to Expert Systems: Knowledge Representation and Problem Solving Methods*. Springer-Verlag, Berlin, 1993.

303. M.L. Puri and D.A. Ralescu. Fuzzy random variables. *Journal of Mathematical Analysis and Applications*, 114:409–422, 1986.

304. J.R. Quinlan. Discovering rules by induction from large collections of examples. In D. Michie, editor, *Expert Systems in the Micro Electronic Age*, pages 168–201. Edinburgh University Press, 1979.

305. J.R. Quinlan. Unknown attribute values in induction. In *Proceedings of the 6th International Workshop on Machine Learning*, pages 164–168, San Mateo, CA, 1989. Morgan Kaufmann.

306. J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

307. R. Quinlan. Combining instance-based and model-based learning. In *Proceedings of the 10th International Conference of Machine Learning*, pages 236–243. Morgan Kaufmann, 1993.

308. A. Ramer. Uniqueness of information measure in the theory of evidence. *Fuzzy Sets and Systems*, 24:183–196, 1987.

309. F.P. Ramsey. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Kegan Paul, London, 1931.

310. C. Reiser and H. Kaindl. Case-based reasoning for multi-step problems and its integration with heuristic search. In J.P. Haton, M. Keane, and M. Manago, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-94, number 984 in LNAI, pages 113–125, Chantilly, France, 1994. Springer-Verlag.

311. F. Ricci and P. Avesani. Learning a local metric for case-based reasoning. In *Proceedings* ICCBR-95, pages 301–312, Sesimbra, Portugal, 1995. Springer-Verlag.

312. M.M. Richter. On the notion of similarity in case-based reasoning. In G.D. Riccia, editor, *Mathematical and Statistical Methods in Artificial Intelligence*, pages 171–184. Springer-Verlag, 1995.

313. M.M. Richter. The knowledge contained in similarity measrues. Invited talk at ICCBR-95. `http://wwwagr.informatik.uni-kl.de/~lsa/cbr/richtericcbr95remarks.html`, 1995.

314. M.M. Richter, S. Wess, K.D. Althoff, and F. Maurer, editors. *First European Workshop on Case-Based Reasoning.* Number 837 in Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 1993.

315. C.K. Riesbeck and R.C. Schank. *Inside Case-based Reasoning.* Hillsdale, New York, 1989.

316. B.D. Ripley. *Spatial Statistics.* Wiley, Chichester, 1981.

317. A.F. Rodriguez, S. Vandera, and L.E. Sucar. A probabilistic model for case-based reasoning. In D.B. Leake and E. Plaza, editors, *Case-Based Reasoning Research and Development, Proceedings of the 2nd International Conference on Case-Based Reasoning,* ICCBR-97, pages 623–632, Providence, RI, USA, 1997. Springer-Verlag.

318. M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

319. A. Rubinstein. Similarity and decision making under risk (is there a utility theory resolution to Allais paradox?). *Journal of Economic Theory*, 46:145–153, 1988.

320. E.H. Ruspini. On the semantics of fuzzy logic. *International Journal of Approximate Reasoning*, 5:45–88, 1991.

321. S.J. Russell. Rationality and intelligence. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 950–957, Montreal, Canada, 1995.

322. S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, New Jersey, 1995.

323. S.J. Russell, D. Subramanian, and R. Parr. Provably bounded optimal agents. In *International Joint Conference on Artificial Intelligence*, pages 338–344, 1993.

324. S.J. Russell and E. Wefald. Principles of metareasoning. *Artificial Intelligence*, 49:361–395, 1991.

325. S.J. Russell and E.H. Wefald. *Do the Right Thing: Studies in Limited Rationality.* MIT Press, Cambridge, Massachusetts, 1991.

326. R. Sabbadin. Decision as abduction. In H. Prade, editor, *Proceedings* ECAI-98*, 13th European Conference on Artificial Intelligence*, pages 600–604, Brighton, UK, 1998.

327. S. Salzberg. A nearest hyperrectangle learning method. *Machine Learning*, 6:251–276, 1991.

328. E. Sanchez. On possibility qualification in natural languages. *Information Sciences*, 15:45–76, 1978.

329. S.A. Sandri, D. Dubois, and K. Kalsfbeek. Elicitation, assessment, and pooling of expert judgements using possibility theory. IEEE *Transactions on Fuzzy Systems*, 3(3):313–335, 1995.

330. R. Sarin and P. Wakker. A simple axiomatization of nonadditive expected utility. *Econometrica*, 60(6):1255–1272, 1992.

331. L.J. Savage. *The Foundations of Statistics*. John Wiley and Sons, Inc., New York, 1954.

332. R. Schank. *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge University Press, New York, 1982.

333. R. Schank and R. Abelson. *Scrips, Goals and Understanding*. Erlbaum, Northvale, NJ, 1977.

334. D. Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57:571–587, 1989. (First Version 1982).

335. B. Schölkopf and AJ. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

336. G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

337. G. Shafer. *The Art of Causal Conjecture*. MIT Press, 1996.

338. C.E. Shannon and W. Weaver. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

339. J. Shawe-Taylor and N. Christianini. *Kernel Methods for Pattern Anylsis*. Cambridge University Press, 2004.

340. D. Shepard. A two-dimensional interpolation function for irregularly spaced data. In *Proceedings of the 23rd National Conference of the ACM*, pages 517–523, 1968.

341. R.N. Shephard. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323, 1987.

342. R. Short and K. Fukunaga. The optimal distance measure for nearest neighbor classification. IEEE *Transactions on Information Theory*, 27:622–627, 1981.

343. B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

344. H.A. Simon. *Models of Man*. John Wiley and Sons, New York, 1957.

345. D.B. Skalak. Prototype and feature selection by sampling and random mutation hill-climbing algorithms. In *Proceedings of the 11th International Conference on Machine Learning*, pages 293–301, New York, 1994. Morgan Kaufmann.

346. D.B. Skalak and E.L. Rissland. Inductive learning in a mixed paradigm setting. In *Proceedings* AAAI-90, pages 840–847, 1990.

347. P. Smets. Information content of an evidence. *International Journal of Man-Machine Studies*, 19:33–43, 1983.

348. P. Smets. Belief functions. In P. Smets, E.H. Mamdani, D. Dubois, and H. Prade, editors, *Non-Standard Logics for Automated Reasoning*, pages 253–277. Academic Press, London, 1988.

349. P. Smets. The concept of distinct evidence. In *Proceedings* IPMU-92, pages 789–794, Palma de Mallorca, Spain, 1992.

350. P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

351. E.E. Smith and D.L. Medin. *Categories and Concepts*. Harvard University Press, Cambridge, MA, 1981.

352. M. Smithon. *Ignorance and Uncertainty*. Springer-Verlag, Berlin, 1989.

353. B. Smyth and P. Cunningham. The utility problem analysed. In J. Smith and B. Faltings, editors, *Advances in Case-Based Reasoning*, pages 392–399. Springer-Verlag, 1996.

354. B. Smyth and M.T. Keane. Adaptation-guided retrieval: questioning the similarity assumption in reasoning. *Artificial Intelligence*, 102(2):249–293, 1998.

355. B. Smyth and T. Keane. Remembering to forget. In C.S. Mellish, editor, *Proceedings International Joint Conference on Artificial Intelligence*, pages 377–382. Morgan Kaufmann, 1995.

356. B. Smyth and E. Mc Kenna. Modelling the competence of case-bases. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning, Proceedings* EWCBR-98, number 1488 in LNAI, pages 208–220. Springer-Verlag, 1998.

357. B. Smyth and E. Mc Kenna. Building compact competent case-bases. In K.D. Althoff, R. Bergmann, and L.K. Branting, editors, *Case-Based Reasoning Research and Development, Proceedings 3rd International Conference on Case-Based Reasoning*, number 1650 in LNAI, pages 329–342, 1999.

358. A. Stahl. Learning of similarity measures: A formal view based on a generalized CBR model. In *Proceedings ICCBR-2005, 6th International Conference on Case-Based Reasoning*, pages 507–521, Chicago, Illinois, 2005. Springer-Verlag.

359. C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, pages 1213–1228, 1986.

360. V. Strassen. Meßfehler und Information. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 2:273–305, 1964.

361. T.M. Strat. Decision analysis using belief functions. *International Journal of Approximate Reasoning*, 4:391–417, 1990.

362. M. Sugeno. *Theory of Fuzzy Integrals and its Application.* PhD thesis, Tokyo Institute of Technology, 1974.

363. J. Surma and J. Tyburcy. A study on competence-preserving case replacing strategies in case-based reasoning. In B. Smyth and P. Cunningham, editors, *Adavances in Case-Based Reasoning, Proceedings* EWCBR-98*, 4th European Workshop on Case-Based Reasoning*, number 1488 in LNAI, pages 233–238. Springer-Verlag, 1998.

364. M. Tan. Cost-sensitive learning of classification knowledge and its application to robotics. *Machine Learning*, 13(7):7–34, 1993.

365. S.W. Tan and J. Pearl. Qualitative decision theory. In *Proceedings* AAAI-94*, 11th National Conference on Artificial Intelligence*, pages 928–933, Seattle, WA, 1994.

366. S.W. Tan and J. Pearl. Specification and evaluation of preferences under uncertainty. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Proceedings* KR-94*, 4th International Conference on Principles of Knowledge Representation and Reasoning*, pages 530–539. Springer-Verlag, Bonn, Germany, 1994.

367. H. Tanaka, S. Uejima, and K. Asai. Linear regression analysis with fuzzy model. IEEE *Transactions on Systems, Man, and Cybernetics*, 12:903–907, 1982.

368. S. Thrun and L. Pratt, editors. *Learning to Learn.* Kluwer Academic Publishers, 1997.

369. H. Tirri, P. Kontkanen, and P. Myllymäki. A Bayesian framework for case-based reasoning. In I. Smith and B. Faltings, editors, *Advances in Case-Based Reasoning, Proceedings* EWCBR-96*, 3rd European Workshop on Case-Based Reasoning*, number 1168 in LNAI, pages 413–427, 1996.

370. I. Tomek. A generalization of the k-NN rule. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-6:121–126, 1976.

371. E. Trillas and L. Valverde. An inquiry in indistinguishability operators. In H.J. Skala et. al., editor, *Aspects of Vagueness*, pages 512–581. Dordrecht, Reidel, 1984.

372. I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proc. ICML–2004*, pages 823–830, Banff, Alberta, Canada, 2004.

373. A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

374. A. Tversky and D. Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

375. A. Tversky and D. Kahneman. Rational choice and the framing of decisions. *Journal of Business*, 59:251–278, 1986.

376. L. Ughetto, D. Dubois, and H. Prade. Implicative and conjunctive fuzzy rules: A tool for reasoning from knowledge and examples. In *Proceedings* AAAI-99, Orlando, 1999.

377. P.E. Utgoff. Incremental induction of decision trees. *Machine Learning*, 4:161–186, 1989.

378. L. Valverde. On the structure of F-indistinguishability operators. *Fuzzy Sets and Systems*, 17:313–328, 1995.

379. V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1982.

380. V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

381. M. Veloso. *Planning and Learning by Analogical Reasoning*. Number 886 in Lecture Notes in Computer Science. Springer-Verlag, Berlin, 1994.

382. M. Veloso and A. Aamodt, editors. ICCBR-95*, First International Conference on Case-Based Reasoning*. Springer-Verlag, Berlin, 1995.

383. C. Wagner and K. Lehrer. *Rational Consensus in Science and Society*. Reidel, Dordrecht, 1981.

384. P. Wakker. A behavioral foundation for fuzzy measures. *Fuzzy Sets and Systems*, 37:327–350, 1990.

385. M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

386. L.X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. IEEE *Transactions on Systems, Man, and Cybernetics*, 22(6):1414–1427, 1992.

387. P.Z. Wang. Treating a fuzzy subset as a projectable random set. In M.M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 213–220. North-Holland, Amsterdam, 1982.

388. Z. Wang and G.J. Klir. *Fuzzy Measure Theory*. Plenum Press, 1992.

389. L.A. Wasserman. Belief functions and statistical inference. *Canadian Journal of Statistics*, 18(3):183–196, 1990.

390. L.A. Wasserman. Prior envelopes based on belief functions. *Annals of Statistics*, 18:454–464, 1990.

391. E.J. Wegmann and I.W. Wright. Splines in statistics. *Journal of the American Statistical Association*, 78:351–365, 1983.

392. K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006.

393. J. Weisbrod. A new approach to fuzzy reasoning. *Soft Computing*, 2:89–99, 1998.

394. S. Wess, K.D. Althoff, and G. Derwand. Using k-d trees to improve the retrieval step in case-based reasoning. In S. Wess, K.D. Althoff, and M.M. Richter, editors, *Topics in Case-Based Reasoning*, pages 167–181. Springer-Verlag, Berlin, 1994.

395. D. Wettschereck and D.W. Aha. Weighting features. In M. Veloso and A. Aamodt, editors, *Case-based reasoning research and development*, pages 347–358, number 1010 in Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 1995.

396. D. Wettschereck, D.W. Aha, and T. Mohri. A review and empirical comparison of feature weighting methods for a class of lazy learning algorithms. *AI Review*, 11:273–314, 1997.

397. D.L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

398. D.R. Wilson. *Advances in Instance-Based Learning Algorithms*. PhD thesis, Department of Computer Science, Brigham Young University, 1997.

399. D.R. Wilson and T.R. Martinez. Improved heterogenious distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

400. IH. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.

401. R.R. Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems*, 9:249–260, 1983.

402. R.R. Yager. Approximate reasoning as a basis for rule-based expert systems. IEEE *Transactions on Systems, Man, and Cybernetics*, 14:636–643, 1984.

403. R.R. Yager. Aggregating evidence using quantified statements. *Information Sciences*, 36:179–206, 1985.

404. R.R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE *Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.

405. R.R. Yager. Decision-making under Dempster-Shafer uncertainties. *International Journal of General Systems*, 20(3):233–245, 1992.

406. R.R. Yager. A general approach to rule aggregation in fuzzy logic control. *Applied Intelligence*, 2(4):333–352, 1992.

407. R.R. Yager. Case-based reasoning, fuzzy systems modelling and solution composition. In D.B. Leake and E. Plaza, editors, *Case-Based Reasoning Research and Development, Proceedings* ICCBR-97, pages 633–643, Providence, RI, USA, 1997. Springer-Verlag.

408. RR. Yager. Prototype based reasoning and fuzzy modeling. In B. Bouchon-Meunier, L. Foulloy, and RR. Yager, editors, *Intelligent Systems for Information Processing*, pages 167–178. Elsevier, 2003.

409. R.R. Yager. Soft aggregation methods in case based reasoning. *Applied Intelligence*, 21:277–288, 2004.

410. M.S. Ying. A logic for approximate reasoning. *The journal of symbolic logic*, 59:830–837, 1994.

411. T.P. Yunck. A technique to identify nearest neighbors. IEEE *Transactions on Systems, Man, and Cybernetics*, 6(10):678–683, 1976.

412. L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

413. L.A. Zadeh. A fuzzy-set theoretic interpretation of linguistic hedges. *J. Cybernetics*, 2(3):4–32, 1972.

414. L.A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-3:28–44, 1973.

415. L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning, parts 1-3. *Information Science*, 8/9, 1975.

416. L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

417. L.A. Zadeh. PRUF: A meaning representation language for natural language. *International Journal of Man-Machine Studies*, 10:395–460, 1978.

418. L.A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 90(2):111–127, 1997.

419. J. Zhang. Selecting typical instances in instance-based learning. In *Proceedings* ICML–92*, 9th International Conference on Machine Learning*, pages 470–479, 1992.

420. J. Zhang, Y. Yim, and J. Yang. Intelligent selection of instances for prediction in lazy learning algorithms. *Artificial Intelligence Review*, 11:175–191, 1997.

421. L.M. Zouhal and T. Denoeux. An evidence-theoretic k-NN rule with parameter optimization. IEEE *Transactions on Systems, Man and Cybernetics – Part C*, 28(2):263–271, 1998.

# Index

admissibility, 84
aggregation operator, 38
alpha-cut, 231
approximation
– case-based, 72

background knowledge, 7, 20
belief, 254, 282
belief function, 132, 226, 277
betting function, 133
bounded optimality, 254

candidate-elimination, 84
case adaptation, 31
case completion, 55
case indexing, 31
case retrieval, 31, 54
case-based inference, 8, 42
CBR cycle, 31
certainty rules, 214, 230
Choquet integral, 39, 277
classification, 32
cluster analysis, 56
cognitive modeling, 30
complete ignorance, 236
concept description, 29
confidence region, 9
consistency principle, 82
cost, 7
cross validation, 179, 214

data analysis, 219
data analysis
– fuzzy, 219
density estimation, 176, 213, 222
density estimation
– kernel-based, 213
Dirac measure, 50
distance, 33

equivalence, 37
equivalence class
– fuzzy, 38
exceptionality, 244, 245
expected utility, 256, 274, 275

frequency, 7, 184, 210, 219, 245
fuzzy graph, 223
fuzzy measure, 129
fuzzy partition, 204
fuzzy quantifier, 270
fuzzy random variables, 222
fuzzy sets, 37

graceful degradation, 54
gradual rules, 230
guaranteed possibility, 169, 173, 192

hyper-parameter, 99
hypothesis space, 17

imperfect specification, 130, 240
implication, 229
implication
– Kleene-Dienes, 236
– Rescher-Gaines, 231, 246
– residuated, 231
– strong, 236
imprecision, 220
induction
– implicit, 20
– rule, 32
inductive bias, 17
inductive learning
– hypothesis, 82
inductive reasoning
– principle, 7
inference
– case-based, 112
– statistical, 106, 112
– transductive, 20
instance level, 43
instance space, 42
interpretative CBR, 32
interval probability, 70

kernel function, 213
kernel smoothing, 56
knowledge revision, 22
knowledge-based systems, 11

learning
– case-based, 41, 81
– incremental, 21
– instance-based, 19, 29
– lazy, 19
– model-based, 19
likelihood principle, 106
linguistic hedges, 200, 201, 213
local-global principle, 35

mass distribution, 132, 143
maximal informativeness
– principle of, 169, 172

maximum likelihood, 18, 71, 221
measurement, 6
meta-knowledge, 47
metric
– discrete, 44
– local, 80
minimal specificity
– principle of, 167, 237
minimum uncertainty principle, 83
mixed uncertainty, 278
modifier, 194, 200, 203, 212
multidimensional scaling, 56

non-specificity, 142

Occam's Razor, 18
outlier, 144
overfitting, 21
OWA operator, 279

parallel computation, 65
plausibility function, 132
possibilistic mixture, 280
possibilistic networks, 222
possibilistic variables, 221
prediction, 8, 41, 55
preference, 6, 254, 282
prior probability, 98
prisoner's dilemma, 271
probabilistic sets, 222
probability
– upper, 153
prototype, 172, 199
proximity relation, 37
pseudometric, 38

random fuzzy sets, 222
random set, 135
rationality, 253
rationality
– bounded, 287
– metalevel, 253
– perfect, 253
reasoning
– knowledge-driven, 13
– analogical, 165
– approximate, 12
– case-based, 30
– data-driven, 13
– exemplar-based, 41
– instance-based, 17
– memory-based, 41
– model-based, 17, 19
– plausible, 12
recognition, 55
regression
– analysis, 119
– linear, 10, 43
– locally weighted, 76
repeated problem solving, 271, 286
representation
– attribute–value, 46
representativeness principle, 284

Shannon entropy, 143
similarity hypothesis, 60
similarity hypothesis
– probabilistic, 108
similarity level, 43
similarity measure, 33
similarity profile, 59
similarity profile
– local, 80
specification context, 240
statistics
– non-parametric, 19
– robust, 56
– semi-parametric, 21
stochastic dominance, 108
structure
– case, 157
– outcome, 77
– similarity, 66
Sugeno integral, 41, 217
symmetry, 34

time series analysis, 120
transferable belief model, 279
transitivity, 34
triangular norm, 37

uncertainty, 6
utility theory, 7

vague data, 219
version space, 84

# THEORY AND DECISION LIBRARY

## SERIES B: MATHEMATICAL AND STATISTICAL METHODS

*Editor*: H. J. Skala, *University of Paderborn, Germany*

1. D. Rasch and M.L. Tiku (eds.): *Robustness of Statistical Methods and Nonparametric Statistics*. 1984 ISBN 90-277-2076-2
2. J.K. Sengupta: *Stochastic Optimization and Economic Models*. 1986 ISBN 90-277-2301-X
3. J. Aczél: *A Short Course on Functional Equations*. Based upon Recent Applications to the Social Behavioral Sciences. 1987 ISBN Hb 90-277-2376-1; Pb 90-277-2377-X
4. J. Kacprzyk and S.A. Orlovski (eds.): *Optimization Models Using Fuzzy Sets and Possibility Theory.* 1987 ISBN 90-277-2492-X
5. A.K. Gupta (ed.): *Advances in Multivariate Statistical Analysis.* Pillai Memorial Volume. 1987 ISBN 90-277-2531-4
6. R. Kruse and K.D. Meyer: *Statistics with Vague Data.* 1987 ISBN 90-277-2562-4
7. J.K. Sengupta: *Applied Mathematics for Economics.* 1987 ISBN 90-277-2588-8
8. H. Bozdogan and A.K. Gupta (eds.): *Multivariate Statistical Modeling and Data Analysis.* 1987 ISBN 90-277-2592-6
9. B.R. Munier (ed.): *Risk, Decision and Rationality.* 1988 ISBN 90-277-2624-8
10. F. Seo and M. Sakawa: *Multiple Criteria Decision Analysis in Regional Planning.* Concepts, Methods and Applications. 1988 ISBN 90-277-2641-8
11. I. Vajda: *Theory of Statistical Inference and Information.* 1989 ISBN 90-277-2781-3
12. J.K. Sengupta: *Efficiency Analysis by Production Frontiers.* The Nonparametric Approach. 1989 ISBN 0-7923-0028-9
13. A. Chikán (ed.): *Progress in Decision, Utility and Risk Theory.* 1991 ISBN 0-7923-1211-2
14. S.E. Rodabaugh, E.P. Klement and U. Höhle (eds.): *Applications of Category Theory to Fuzzy Subsets*. 1992 ISBN 0-7923-1511-1
15. A. Rapoport: *Decision Theory and Decision Behaviour.* Normative and Descriptive Approaches. 1989 ISBN 0-7923-0297-4
16. A. Chikán (ed.): *Inventory Models.* 1990 ISBN 0-7923-0494-2
17. T. Bromek and E. Pleszczyńska (eds.): *Statistical Inference.* Theory and Practice. 1991 ISBN 0-7923-0718-6
18. J. Kacprzyk and M. Fedrizzi (eds.): *Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory.* 1990 ISBN 0-7923-0884-0
19. G.L. Gómez M.: *Dynamic Probabilistic Models and Social Structure.* Essays on Socioeconomic Continuity. 1992 ISBN 0-7923-1713-0
20. H. Bandemer and W. Näther: *Fuzzy Data Analysis.* 1992 ISBN 0-7923-1772-6
21. A.G. Sukharev: *Minimax Models in the Theory of Numerical Methods.* 1992 ISBN 0-7923-1821-8
22. J. Geweke (ed.): *Decision Making under Risk and Uncertainty.* New Models and Empirical Findings. 1992 ISBN 0-7923-1904-4
23. T. Kariya: *Quantitative Methods for Portfolio Analysis*. MTV Model Approach. 1993 ISBN 0-7923-2254-1
24. M.J. Panik: *Fundamentals of Convex Analysis*. Duality, Separation, Representation, and Resolution. 1993 ISBN 0-7923-2279-7

# THEORY AND DECISION LIBRARY: SERIES B

25. J.K. Sengupta: *Econometrics of Information and Efficiency.* 1993 ISBN 0-7923-2353-X
26. B.R. Munier (ed.): *Markets, Risk and Money*. Essays in Honor of Maurice Allais. 1995 ISBN 0-7923-2578-8
27. D. Denneberg: *Non-Additive Measure and Integral*. 1994 ISBN 0-7923-2840-X
28. V.L. Girko, *Statistical Analysis of Observations of Increasing Dimension*. 1995 ISBN 0-7923-2886-8
29. B.R. Munier and M.J. Machina (eds.): *Models and Experiments in Risk and Rationality*. 1994 ISBN 0-7923-3031-5
30. M. Grabisch, H.T. Nguyen and E.A. Walker: *Fundamentals of Uncertainty Calculi with Applications to Fuzzy Inference*. 1995 ISBN 0-7923-3175-3
31. D. Helbing: *Quantitative Sociodynamics*. Stochastic Methods and Models of Social Interaction Processes. 1995 ISBN 0-7923-3192-3
32. U. Höhle and E.P. Klement (eds.): *Non-Classical Logics and Their Applications to Fuzzy Subsets*. A Handbook of the Mathematical Foundations of Fuzzy Set Theory. 1995 ISBN 0-7923-3194-X
33. M. Wygralak: *Vaguely Defined Objects.* Representations, Fuzzy Sets and Nonclassical Cardinality Theory. 1996 ISBN 0-7923-3850-2
34. D. Bosq and H.T. Nguyen: *A Course in Stochastic Processes.* Stochastic Models and Statistical Inference. 1996 ISBN 0-7923-4087-6
35. R. Nau, E. Grønn, M. Machina and O. Bergland (eds.): *Economic and Environmental Risk and Uncertainty*. New Models and Methods. 1997 ISBN 0-7923-4556-8
36. M. Pirlot and Ph. Vincke: *Semiorders.* Properties, Representations, Applications. 1997 ISBN 0-7923-4617-3
37. I.R. Goodman, R.P.S. Mahler and H.T. Nguyen: *Mathematics of Data Fusion*. 1997 ISBN 0-7923-4674-2
38. H.T. Nguyen and V. Kreinovich: *Applications of Continuous Mathematics to Computer Science.* 1997 ISBN 0-7923-4722-6
39. F. Aleskerov: *Arrovian Aggregation Model*. 1999 ISBN 0-7923-8451-2
40. M.J. Machina and B. Munier (eds.): *Beliefs, Interactions and Preferences in Decision Making*. 1999 ISBN 0-7923-8599-3
41. V. Serdobolskii: *Multivariate Statistical Analysis.* A High-Dimensional Approach. 2000 ISBN 0-7923-6643-3
42. A. Gore and S. Paranjpe: *A Course in Mathematical and Statistical Ecology.* 2001 ISBN 0-7923-6715-4
43. S. Li, Y. Ogura, V. Kreinovich: *Limit Theorems and Applications of Set-Valued and Fuzzy Set-Valued Random Variables*. 2002 ISBN 1-4020-0918-6
44. E. Hüllermeier: *Case-Based Approximate Reasoning*. 2007 ISBN 1-4020-5694-X

springer.com