Probabilistic Interpretations
of
Fuzzy Sets and Systems

by
Ying Li

B.S., Tsing Hua University (1986)
M.S., Massachusetts Institute of Technology (1989)

Submitted to the Department of Electrical Engineering and
Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

July 1994

© Massachusetts Institute of Technology 1994. All rights reserved.

Author .................................................................
Department of Electrical Engineering and Computer Science
July 20, 1994

Certified by .........................................................
Robert G. Gallager
Fujitsu Professor of Electrical Engineering
Thesis Supervisor

Accepted by .........................................................
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students
Probabilistic Interpretations

of

Fuzzy Sets and Systems

by

Ying Li

Submitted to the Department of Electrical Engineering and Computer Science on July 20, 1994, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Fuzzy sets are generalized sets for which each element is in the set to some degree, as opposed to ordinary sets in which each element is either completely in or not in the set. Proposed as a tool for modeling imprecise classes such as "small numbers", fuzzy sets have been studied in many research areas and widely used in industrial processes and consumer products. The strength of using fuzzy sets to model imprecise classes is the ability to use linguistic descriptions and human insights in modeling. Despite their popularity, fuzzy sets contain difficulties; for example, statements such as "If A then not A" can be true in fuzzy set theory, even when A is not empty. There are also various ad hoc operations for taking union and intersection. Many people have questioned conventional fuzzy set theory and proposed probability as an alternative. In this report, we explore the probabilistic interpretations of fuzzy sets systematically, and show how the above difficulties get resolved. We discuss different ways to construct conceptual probabilistic experiments, and their common features. The idea of an enlarged sample space, which can be viewed as the cartesian product of an observation space and a description space, is introduced. Fuzzy sets are viewed as events (crisp sets) in the enlarged sample space. The unified language allows established results in probability to be used for fuzzy sets, and concepts in fuzzy set theory to be shared with problems usually solved by probability. In fuzzy control, probability provides guidelines in choosing membership functions and operations. In fuzzy pattern recognition, probability suggests ways to better use contextual information to improve classification accuracy. We conclude that the mathematical structures of probability can be used to the benefit of fuzzy set theory.

Thesis Supervisor: Robert G. Gallager
Title: Fujitsu Professor of Electrical Engineering
Acknowledgments

I truly appreciate Professor Robert G. Gallager. As my thesis advisor, his insights and guidance have been most valuable. I thank him for his advice, support, compassion, and humor over the years. The lectures he gave are full of intuition and highly enjoyable. Those who do not have the good fortune of attending his classes should not miss his books. A new one is coming out, “Discrete Stochastic Processes”.

Professor Alvin Drake and Professor Sanjoy Mitter, my readers, provided lot’s of support and many helpful suggestions regarding this thesis. Working as a teaching assistant in Professor Drake’s probability course is just about the most interesting teaching and learning experience. The discussions and encouragements from Professor Mitter about this research are deeply appreciated.

Special thanks go to Dr. H. Witsenhausen, who has been most generous in sharing his wisdom. Professor C. Searle offered me my first teaching assistantship. Dr. D. G. Forney wrote the reference letter in the shortest possible time. Professor V. Bier offered much help in starting my study of fuzzy set theory. Discussions with Prof. L. X. Wang, Dr. J. S. Jang, Prof. M. Trott and Prof. D. Bertesekas are much appreciated. I am also grateful for the support from Professor C. F. Chen.

The many talented members (or former members) of LIDS are all truly remarkable, and their friendship will always be treasured. I thank Jay Kuo, Whay Lee, John Spinelli, Emre Telatar, Rajesh Pankaj, Jane Simmons, Abhay Parekh, Roy Yates, Murat Azizoglu, Mike Chin, Daniel Lee, Jinane Abunadi, Diane Ho, Muriel Medard, David Tse, Jen Chen, Serap Sevari, May Ku, Diana Debby, Angela Chiu, Li-Chung Chang, Rachael Learned, Seema Jaggi, Peter Li, Stefano Casadei, and Mike Daniel. I thank Mohamad Akra, Mike Branicky and Wesley McDermott for discussions of their research. I thank Nancy, Kathleen, Shiel, Fifa, and Margaret for their help and good humor. I thank Marilyn, Peggy, Monica, and Lisa Bella in the EECS headquarters for their concern and for being so helpful.

I am blessed with many wonderful friends. I thank Yuan Junn, Li Su, Abraham, Jin Lien, Cynthia, Szu Li, Meng Yong, Mark, Jocelyn, Eric, Mike, Phillip, Kelly, Ting
Jen, Ian, Hsien Jen, Ya Hwei, Jen Fong, Su Jen, Hen Ru, Ping, Cheh Chih, Hong Tat, Yu Hwei and Yu Kang.

I have the most lovely host, Ms. Rosallie Allen-Ryan. I thank her for her hospitality and friendship.

Very special thanks go to a most unusual friend, Nate Osgood. In addition to being extremely supportive, Nate has proof read my thesis and provided many excellent suggestions.

I am deeply indebted to my family. Many thanks and much love go to my parents, Ke-Yang Li and Ching-Wen Hsieh. I also thank my brothers and sisters, Yen Jay, Lee Yuan, Wan Ru and Hsiao Chiang. The courage, love, and support—particularly in the past two years—from my mother-in-law Yuan Yang, my husband Chin Hsun, and my son Jeffery is crucial for the completion of this thesis. No words can describe my appreciation towards them.

This research was partially supported by the Vinton Hayes fellowship, the Grass Instrument Company fellowship, funds from the Fujitsu Chair of Electrical Engineering, and ARO grant DAAL03-92-G-0115. I am thankful for the funding.
# Contents

1 Introduction

1.1 Fuzzy Sets: an Overview ........................................ 7
1.2 Issues in Fuzzy Set Theory .................................... 9
1.3 Fuzzy Set Theory and Probability ............................ 10
1.4 Issues in Probability ........................................... 11
1.5 Thesis Outline .................................................. 14

2 Basic Concepts of Fuzzy Sets ................................. 16

2.1 Why Fuzzy Sets? ............................................... 16
2.2 Fuzzy Set Definitions ......................................... 18
  2.2.1 Membership Function .................................... 18
  2.2.2 Combining Fuzzy Sets .................................... 19
  2.2.3 Relations Between Fuzzy Sets .......................... 21
  2.2.4 Issues of Concern ......................................... 21

3 Probabilistic Interpretations of Fuzzy Sets .................. 25

3.1 Introduction .................................................. 25
3.2 Borel's Idea: Polling a Population .......................... 27
  3.2.1 Describing a Heap of Wheat ............................. 27
  3.2.2 The Probabilistic Interpretation of Borel's Idea  ....... 27
  3.2.3 Finding the Membership Function as a Fraction of a Population 29
3.3 Mitter's Idea: Random Intervals .............................. 30
3.4 The Probabilistic Model for Fuzzy Sets ....................... 34
3.4.1 Modeling Fuzzy Sets as Events in the Enlarged Sample Space 
3.4.2 General Structure of the Enlarged Sample Space 
3.4.3 Visualization of the Enlarged Sample Space 
3.4.4 Finding Membership Functions for Multiple Fuzzy Sets 
3.4.5 Fuzzy Interpretations of Random Phenomena 
3.4.6 Viewing Fuzzy Sets as Events (Sets), Not Functions 
3.5 Relations between Fuzzy Sets: Our View 
3.5.1 The Basic Relations 
3.5.2 The Choice of Fuzzy Union/Intersection Operations 
3.5.3 Conditional Independence of Fuzzy Sets Describing Different Observations 
3.6 The Pseudomembership Function 
3.6.1 What is a Pseudomembership Function and When Should It be Used 
3.6.2 Quantization-type Fuzzy Sets and Detection-type Fuzzy Sets 
3.6.3 Conditional Independence of Observations Given Fuzzy Sets Describing Them 
3.7 Common Relations for Fuzzy Sets Describing Multiple Variables 
3.8 Probability and Fuzzy Sets: Zadeh's View, Our View 
3.9 Conclusions 

4 Fuzzy Control 
4.1 Introduction 
4.2 Fuzzy control in Control Engineering 
4.3 Structure of a Fuzzy Controller 
4.3.1 Pure Fuzzy Controllers 
4.3.2 Sugeno Fuzzy Controllers 
4.4 Probabilistic Interpretations of a Fuzzy Controller 
4.4.1 Pure Fuzzy Controllers 
4.4.2 Sugeno Fuzzy Controllers
List of Figures

2-1 “Small” as a crisp set and as a fuzzy set. ........................................ 17
2-2 Membership function of fuzzy set “numbers close to r”. .................. 19
2-3 Three fuzzy sets that form a fuzzy partition on [0, 10] ....................... 22
3-1 Triangular Fuzzy Set ................................................................. 32
3-2 Trapezoidal Fuzzy Sets............................................................. 32
3-3 Dithered Quantization. ............................................................. 33
3-4 Enlarge the observation space to obtain the sample space. ............... 36
3-5 Fuzzy sets are events (crisp sets) in the enlarged sample space. ....... 37
3-6 Sample space as the Cartesian product of observation space and de-
  scription space. ........................................................................... 39
3-7 Visualize the enlarged sample space: discrete observation space. ..... 41
3-8 Visualize the enlarged sample space: observation space continuous... 43
3-9 The sample space in Borel’s probabilistic experiment. .................... 44
3-10 Probabilities in the binary communication problem. ....................... 46
3-11 Fuzzy sets viewed as sets in the enlarged space. ............................ 47
3-12 Two disjoint fuzzy sets. ............................................................ 49
3-13 Inclusion (a) and conditional inclusion (b). ................................... 50
3-14 $m_A(x) \leq m_B(x) \forall x$, but $A$ is not a subset of $B$. .................... 51
3-15 Two fuzzy sets with the same membership functions that are indepen-
  dent conditional on the observation. $P(AB|x) = P(A|x)P(B|x)$. ........ 52
3-16 Two fuzzy sets with the same membership functions that are indepen-
  dent unconditionally. $P(AB) = P(A)P(B)$. ................................. 53

8
3-17 Three fuzzy sets (events) that form a fuzzy partition (partition the sample space). .......................................................... 54
3-18 Three fuzzy sets whose membership functions add up to one. .......... 56
3-19 Fuzzy Sets "Small" and "Large" on $U = [0, 1]$. .......................... 56

4-1 General structure of a control system. ........................................ 69
4-2 General structure of a fuzzy controller. ...................................... 75
4-3 Pure Fuzzy Controller: a Probability Tree Representation. ......... 81
4-4 Probabilistic View of a Fuzzy Controller. ................................. 82
4-5 Equivalent Input Membership functions. ................................. 89
4-6 Gaussian-type fuzzy membership functions. ............................... 90
4-7 Cascading Two Fuzzy Controllers. ........................................... 94
4-8 Fuzzy sets with triangular membership functions. ....................... 99
4-9 A fuzzy controller as (a) a piecewise linear approximator, (b) a piecewise second order polynomial approximator. .................. 101

5-1 Recognition as Searching for Patterns. .................................... 116
5-2 The reduced single sided Hausdorff distance is not perfect. ........ 120
5-3 The hierarchical structure of text recognition. ........................... 123

B-1 A conventional probabilistic model of the character generation process. 134
Chapter 1

Introduction

Fuzzy sets, introduced by Zadeh in 1965 [Zad 65], are generalized sets where each element can be in the set to some degree. Generalizing sets into fuzzy sets is a very natural idea and is widely accepted. However, the precise interpretation of fuzzy sets, and the construction of fuzzy sets from other fuzzy sets by union or intersection, for example, have long been the subject of debates.

We believe that probability can be used as the mathematical structure for fuzzy set theory, as an alternative to the mathematical structure employed in conventional fuzzy set theory. This thesis aims to explore this in a systematic way. We start from the basic picture of the enlarged sample space, view fuzzy sets as events, and study the changes brought by probabilistic interpretations to fuzzy sets and systems.

1.1 Fuzzy Sets: an Overview

In classical set theory, an element is either in or not in a set. A fuzzy set allows each element to belong to the set to some degree. This makes possible gradual transitions from being in the set to not in the set, which is more natural for many real world classes. Fuzzy sets are proposed to better model these imprecise classes. These classes can occur in nature, in natural language and human thinking, or in any system where
imprecision is to be exploited. 1.

Most of the foundation of fuzzy set theory was laid by Zadeh, starting with his paper in 1965 [Zad 65]. Today, fuzzy sets appear in a wide range of areas, from purely theoretic fuzzy mathematics research to the implementation of fuzzy control in various industrial processes and commercial products. Japan was the first country in which fuzzy control was brought into extensive commercial use. Fuzzy logic chips are used in air conditioners, video cameras, washing machines, elevators, automobiles, just to name a few. More than 50 Japanese companies are developing fuzzy applications and chips. In North America, the market for fuzzy controllers was on the order of 2 million in '91 [MiW 91]. U. S. companies and research institutes are also starting to take a serious interest in fuzzy systems.

In academia, journals for fuzzy sets and systems include the IEEE Transaction on Fuzzy Systems, Fuzzy Sets and Systems (North Holland), the International Journal of Approximate Reasoning (Elsevier), the SOFT Journal (Japan), and Fuzzy Mathematics (China). There can well be more than 1000 papers in fuzzy set theory and applications being published yearly.

Despite the flourishing interests, there has been long lasting skepticism towards fuzzy set theory. The most recent round of debates appear in the February 1994 issue of IEEE Transactions on Fuzzy Systems, which is a special issue on Fuzziness Vs. Probability.

While some researchers hold the view that fuzzy sets model nonstatistical uncertainty and can not be dealt with using probability [Zad 86][Bez 94][Kli 94][Kos 92], other researchers argue that probability not only can be used to deal with fuzzy imprecision, but also does it better [Lin 87][Che 86] [His 86][La5 94]. Most researchers agree that it is important to explore both fuzziness and randomness [Gai 78] [His 86][DuP 94].

---

1By “imprecision” we are referring to concepts such as “large” “small” that are less precise than the numerical values they describe. Note that “small” is imprecise not only because it refers to multiple numerical values (knowing “x is small” does not tell us the numerical value of x), but also because a number can be small “to some degree” and not small “to some degree” (knowing x = .3 does not tell us precisely whether x is small or x is not small).
Our views about the debates are as follows:

In the most general sense, fuzzy set theory is the theory that deals with generalized sets where members can be in the set to some degree. Conventional fuzzy set theory has played an important role in exploring imprecision, a concept which has not been systematically distinguished from randomness in probability before the advent of fuzzy set theory\(^2\).

A major issue in the debates surrounding fuzzy set theory is the generalization of set theoretic operations such as union or intersection for use with fuzzy sets. We think that probability provides a flexible way to resolve this issue. Adopting the mathematical structure of probability does not lessen the major contributions of conventional fuzzy set theory. Fuzzy set theory and probability have much to offer one another. Incorporating the mathematical structure of probability into fuzzy set theory will add new dimensions to both theories.

1.2 Issues in Fuzzy Set Theory

Issues often raised against fuzzy set theory are as follows.

- Practical Issues: Basic operations such as union and intersection are often selected in an ad hoc way. Zadeh originally defined $\max/\min$ as the union/intersection operations. Bellman and Gertz [BeZ 70] showed that $\max/\min$ are the only operations that satisfy a set of reasonable rules. In practice $\max/\min$ do not always work well, however. Many other operations have subsequently been proposed, but there are essentially no guidelines for choosing among them. Researchers occasionally need to create their own customized operations to make things work[ZBD 87].

- Theoretical Issues: Fuzzy sets do not satisfy basic laws of logic such as the law of contradiction (the intersection of a fuzzy set and its complement is not

\(^2\)The fuzzy concept “how many grains constitute a heap” was addressed probabilistically by Borel in 1914, who is perhaps the first to to propose the idea of polling a population (see chapter 3).
always empty), and the law of excluded middle (the union of a fuzzy set and its complement is not always the universe). By varying the definition of negation and the union/intersection operations, the law of excluded middle can be satisfied, but only at the cost of violating idempotency and distributivity [FoK 88]. Many fuzzy set theorists consider these violations acceptable, while others view them as weaknesses of fuzzy set theory.

1.3 Fuzzy Set Theory and Probability

Fuzzy set theory deals with fuzziness or imprecision. Concepts such as "small", "large" are imprecise (fuzzy) since a number can be small to some degree and not small to some degree, instead of being precisely small or not small \(^3\).

Probability usually deals with randomness. There is an experiment with multiple possible outcomes, each of which has a given probability. When the experiment is performed, one and only one outcome occurs. An example is coin flipping. The outcome is either a head with one probability or a tail with another probability.

Randomness and fuzziness are different types of uncertainty ("whether" versus "how much" [Kos 92]). However, we believe that probability can also interpret fuzziness through the construction of a conceptual probabilistic experiment.

The idea that fuzziness can be modeled by probability is not new. As early as in 1914, Borel [Bor 14] has described the fuzzy concept "a heap of wheat" with a population idea that is essentially probabilistic. More recent works in this area include comparisons between fuzzy and probabilistic approaches[Gai 78][Sta 77][Bie 92], discussions about why probability is better in dealing with imprecision [Lin 87] [LaS 94], and the probabilistic interpretation of fuzzy membership functions [Che 86][His 86]. There are also mixed opinions or complete rejection of the idea that fuzzy sets can be modeled by probability, [Zad 86] [DuP 94] [Kli 94] [Kos 94] [Wil 94].

In this thesis we study the probabilistic interpretations of fuzzy sets, and the im-

\(^3\) Another aspect of the imprecision of "small" is that there are many numbers that are small (to some degree).
pact to fuzzy control and fuzzy pattern recognition systems. One contribution of this thesis is to clarify the idea of the enlarged sample space, which has not received much attention in previous works. A clear picture of the sample space—the foundation of a probabilistic model—is essential in understanding the probabilistic interpretations of fuzzy sets.

Using probability as the mathematical structure of fuzzy set theory allows fuzzy sets to satisfy all the basic laws of logic, including the law of excluded middle. Probability has the additional structure that set theory lacks for constructing a flexible fuzzy set theory. In our probabilistic model, fuzzy sets become events (crisp sets) in an enlarged underlying sample space. Relations between fuzzy sets become relations between events, and union/intersection are taken according to these relations. The many established results in probability can be applied to fuzzy sets. On the other hand, concepts in fuzzy set theory can be translated into probability terms and be applied to problems dealing with randomness.

1.4 Issues in Probability

This section addresses some issues of concern regarding the application of probability to fuzzy sets.

Fuzzy imprecision is different from randomness. Can probability be used to interpret fuzziness?

For example, “This apple is half small and half large” is clearly not the same as “This apple is either small or large with probability one half” [Kos 92]. One way to model fuzziness like this with probability is to assume that there is a population of people capable of making yes/no decisions. When presented with the half large half small apple and asked the question “Is this apple small?” the fraction of the population that answer “yes” can be taken as the degree of truth that the apple is small. This fraction corresponds to the probability that a randomly chosen person from the population answers “yes” to the above question. This idea is due to Borel
Borel's population idea provides a standard way to interpret fuzzy concepts probabilistically. We need to imagine that there exists a yes/no-decision-making population, and a person is randomly selected to answer the question. Alternatively we might imagine this population to exist in an expert's mind, in which case the expert is making random decisions (see Ch. 3 for more detail). For each of these methods, a conceptual probabilistic experiment is constructed. Unlike the experiment of flipping a coin, in which one event occurs with certain probability, these conceptual experiments for modeling fuzziness are usually not actually performed. This does not disqualify probability as a useful modeling tool.

Whether or not a model is useful depends on whether it provides insights into the system being modeled. Whether the physical process in the system occurs in the same way as in the model is a separate issue and is not required for the model to be useful.

For example, we often model a two terminal linear circuit by its Thevenin equivalent (a voltage source in series with a resistor) or its Norton equivalent (a current source in parallel with a resistor). They are equally good as modeling tools even though there is a current flowing in the Norton equivalent, but not in the Thevenin (an open circuit). The actual circuit being modeled might look totally different from the equivalent circuits (it can contain many more components and connections). None of these affect the usefulness of the Thevenin and Norton equivalent circuits as modeling tools.

The reason why we use probability is not because problems such as "Is this apple small?" is inherently random. We can model the situation where the apple is half small and half large without requiring that either small or large appear as the result of the experiment; the mathematics is independent of that.
Why further define probability for the elements of fuzzy sets, even though it is not necessary?

To interpret fuzzy sets probabilistically, we model the membership degree of an element \( z \) in a fuzzy set \( A \) as a probability \( P(A|z) \). Although it is not required, we will in addition define a probability for each element, \( P_X(x) \), much like the Bayesian approach (define probability for everything). Even in situations where one does not want to (or cannot) assign probability for the elements, there is still a reason for assuming everything has a probability initially, as discussed below.

Assigning everything a probability makes the mathematical model complete. Many problems are more easily solved by adding an extra piece to complete the picture. It is a common technique for obtaining nice non-Bayesian results to define a prior probability that is left unspecified in the beginning, then optimize over all possible prior probabilities to remove it. For example, a communication channel with input \( X \) and output \( Y \) is completely described by the transition probability \( P_{Y|X}(y|x) \). The channel capacity is derived by first assuming there is an unknown input probability \( P_X \), then optimize over \( P_X \). Another example is the Neyman Pearson test. The test only involves the transition probabilities, but it can be derived much more easily if an unspecified input probability is initially assumed.

How about additional complexity?

With a probabilistic interpretation, it is not sufficient, when there are multiple fuzzy sets, to specify each fuzzy set individually. This can be dealt with by assuming simple relations between the fuzzy sets, and is discussed in Chapter 3. The added structure actually provides more choices for simplification.

In terms of the impact to applications, we study fuzzy control and fuzzy pattern recognition. Fuzzy set theory brings to these fields concepts such as linguistic descriptions, interpolative reasoning, soft decisions, and gradual boundaries. We study how these concepts can be captured using the probabilistic interpretations of fuzzy sets.
1.5 Thesis Outline

This thesis is organized as follows:

Chapter 2: Basic Concepts of Fuzzy Sets

Some basic definitions of existing fuzzy set theory are introduced. Topics covered include the membership functions, union and intersection operations, and relations between fuzzy sets. Issues of concern in fuzzy set theory are discussed again in more detail.

Chapter 3: Probabilistic Interpretations of Fuzzy Sets

The general way to model fuzzy sets probabilistically is presented. Borel's population idea and Mitter's random interval idea \(^4\) are discussed first. We then point out that the key idea is to enlarge the sample space, and suggest that the sample space can be viewed as the Cartesian product of an observation space and a description space. Fuzzy sets become events (crisp sets) in the enlarged sample space, and membership functions are conditional probabilities. New ways to visualize fuzzy sets are introduced. We discuss relations between fuzzy sets, and present the idea of pseudomembership functions. We compare our results with the results of Zadeh given in [Zad 68].

Chapter 4: Fuzzy Control

A summary of Tomizuka's comments on general issues about fuzzy control is presented first. Basic structures of fuzzy controllers are first studied in the conventional view, then in the probabilistic view. Probability suggests the use of normalized fuzzy sets, product fuzzification, product inference, and sum composition. We also study fuzzy controllers as function approximators. The rate of convergence in terms of the

\(^4\)This idea was discussed independently by other researchers and was referred to as the random set idea in [DuP 94].
number of fuzzy if-then rules and degrees of input/output fuzzy membership functions is computed. The trade-off between simple linguistic descriptions and numerical efficiency is discussed.

Chapter 5: Fuzzy Pattern Recognition.

We study fuzzy pattern recognition where the final decision is the most likely class (a crisp or hard decision). Fuzzy pattern recognition is viewed as a soft decision approach. It is compared to probabilistic soft decoding, an application of Bayesian detection. When the candidate classes are modeled as fuzzy sets that partition the enlarged sample space, they can be viewed as underlying hypotheses, and fuzzy pattern recognition can be carried out as Bayesian detection. This adds an improved ability to use contextual information to fuzzy pattern recognition. We apply fuzzy pattern recognition based on Bayesian detection to Akra’s [Akr 93] automated text recognition scheme to provide soft decisions and gradual boundaries.
Chapter 2

Basic Concepts of Fuzzy Sets

2.1 Why Fuzzy Sets?

Many natural classes are fuzzy in the sense that the transition from being in the class to not in the class is gradual rather than abrupt [Zad 65]. These classes are appropriately modeled by fuzzy sets, where each element can be in the set "to some degree".

Consider the concept small. If one was presented with an item and asked: "Is this item small?", it might be difficult to give a clear yes/no answer. There are different kinds of ambiguities. The answer depends on the object we are talking about (e.g. "small rock", "small town"...), the observer (e.g. what seems to be small to an adult might not be so to a child), the time (e.g. a dollar is a small amount today, but not so fifty years ago). Without knowing the context in which the question was asked, one cannot be sure of a reasonable answer. The ambiguity here is due to ignorance of context.

On the other hand, even if the context is completely specified, there are still things of medium size and are thus "sort of" small, making it difficult to give a yes/no answer. The ambiguity here is due to the fuzziness of the concept small. To see this better, take for example the set of small numbers in the unit interval $[0, 1]$. If small is modeled as a crisp set, it might contain say numbers less than .5 (see Fig. 2.1(a)). Some undesirable things occur:
• .4999 is small but .5001 is not small even though the difference between them is tiny. It is unnatural to draw a cutting line separating small and not-small numbers.

• 0.0001 and .4999 are both small, the fact that 0.0001 is smaller is not accounted for.

![Diagram](image)

Figure 2-1: “Small” as a crisp set and as a fuzzy set.

These can be solved by modeling small as a fuzzy set, and associating each number in [0, 1] with a **degree of being small**. This degree can decrease smoothly as the number grows (Fig. 2.1(b)). The number .5001 will have a slightly smaller degree of being small than .4999. There will no longer be an abrupt change in the degree of being small. 0.0001 will have a much higher degree of being small than .4999 does.

Why are we interested in fuzzy concepts? Instead of dealing with numerical values, human beings often use fuzzy linguistic descriptions such as small, large in everyday life. Human reasoning is imprecise in nature. Precision is costly. It inhibits thinking by forcing people to focus on detail. It pays to exploit imprecision. An example Zadeh used is the task of parking a car. It would be very difficult if the parking location is specified to within one hundredth of an inch. When the requirement on precision is loosened, the task is greatly simplified. Many people can park the car successfully with simple commands like “turn right a little”. Crisp sets do not capture well such concepts as “small”. Fuzzy sets can provide better associations between linguistic
descriptions and numerical values. Fuzzy sets are better tools for dealing with fuzzy imprecision than crisp sets.

Describing something with fuzzy sets might not imply "reduced precision" in the expected way. The "degree of being small" is a real number. We are not really reducing the amount of precision while describing a numerical value $x$ (a real number) by its degree of being small $m_{small}(x)$ (another real number). However, $m_{small}(x)$ will not be used in the same way as $x$. The information about $x$ is in a different form now and can be more convenient to utilize. It is divided into the linguistic aspect (small) and the numerical aspect ($m_{small}(x)$). Often we can reason on a larger scale in terms of the less precise linguistic descriptions only, without worrying about the specific values of $x$ and $m_{small}(x)$. This can greatly simplify the task. The numerical aspect comes into the picture only when it is necessary to be more precise or to convert between linguistic descriptions and numerical values $x$. Fuzzy control is a good example (see Chapter 4).

### 2.2 Fuzzy Set Definitions

#### 2.2.1 Membership Function

A fuzzy set $A$ defined over a collection of elements $U$ ($U$ is usually called "the universe of discourse") is completely characterized by its membership function $m_A : U \rightarrow [0,1]$. The membership function associates with each element $x$ of $U$ a number $m_A(x)$ in the interval $[0,1]$ that represents the grade of membership of $x$ in $A$. When $A$ is an ordinary set (crisp set), its membership function can take on only two values 1 and 0, with $m_A(x) = 1$ or 0 depending on whether $x$ does or does not belong to $A$.

**Example** Let $U$ be $R$, the real numbers, and $A$ be the fuzzy set of numbers close to a particular number $r$. The membership function $m_A(x)$ might be (Fig. 2.2):

Note that the above is just one out of many possible choices of $m_A(x)$. For the fuzzy set "numbers close to $r" it seems reasonable for the membership function to have the
Figure 2-2: Membership function of fuzzy set “numbers close to \( r \)."

form

\[ m_A(x) = g(d(x, r)), \]

where \( d(x, r) \) is the distance from \( x \) to \( r \), and \( g \) is a nonincreasing function from \([0, \infty)\) onto \([0, 1]\) that approaches zero with increasing distance. Given this, there are still infinitely many possible choices of \( g \). Unlike in probability, where a central limit theorem may sometimes justify the choice of a Gaussian density, there is no similar rule in choosing membership functions.

There is an isomorphism between set theory and binary logic. Every result in one theory has a counterpart in the other theory. Similarly, there is also an isomorphism between fuzzy set theory and fuzzy logic [FoK 88]. The membership function value \( m_A(x) \) in fuzzy set theory corresponds to the truth value of proposition “\( x \) is in \( A \)” in fuzzy logic. \( \text{Union/intersection} \) correspond to \( \text{OR/AND} \), \( A \subseteq B \) corresponds to \( A \rightarrow B \), and so on.

2.2.2 Combining Fuzzy Sets

The most basic ways to combine fuzzy sets are by taking unions and intersections. The following definitions are due to Zadeh [Zad 65].

- **union** \( A \cup B \leftrightarrow m_{A \cup B}(x) = \max(m_A(x), m_B(x)) \).

- **intersection** \( A \cap B \leftrightarrow m_{A \cap B}(x) = \min(m_A(x), m_B(x)) \).
There are various other operations for taking union and intersection. See [FoK 88] [DuP 80]. Any particular intersection operation can be specified by a function $f_i$:

$$f_i : [0,1] \times [0,1] \rightarrow [0,1]$$

$$m_{A \cap B}(x) = f_i[m_A(x), m_B(x)].$$

Similarly, any particular union operation can be specified by a function $f_u$:

$$f_u : [0,1] \times [0,1] \rightarrow [0,1]$$

$$m_{A \cup B}(x) = f_u[m_A(x), m_B(x)].$$

Many alternative fuzzy union/intersection operations are parameterized and are considerably more complicated than $max/min$. Consider the following suggested in [DuP 90]. Let $a = m_A(x)$, $b = m_B(x)$. $\alpha \in (0,1)$.

$$m_{A \cup B}(x) = \frac{a + b - ab - \min(a, b, 1 - \alpha)}{\max(1 - a, 1 - b, \alpha)}$$

$$m_{A \cap B}(x) = \frac{ab}{\max(a, b, \alpha)}$$

Often, as we shall see, it is desirable for the operation $f_i$ or $f_u$ to depend on the particular fuzzy sets $A$ and $B$. In Zadeh’s original theory, the same union or intersection operation is used to combine any two fuzzy sets. This is due to an assumption in fuzzy set theory, sometimes called “truth functionality” in the literature. “Truth functionality” is the property that the truth values of composites are completely determined by truth values of the components. For reasons to be brought out later, we do not believe that this is an essential characteristic of fuzzy set theory.
2.2.3 Relations Between Fuzzy Sets

Let \( A \) and \( B \) be any two fuzzy sets defined on \( x \in U \). In the basic theory developed by Zadeh, relations \(^1\) between any fuzzy sets are completely determined by their individual membership functions. This is again due to “truth functionality”.

- \( A \) and \( B \) are said to be equal \( (A = B) \) when
  \[
  m_A(x) = m_B(x) \quad \forall x \in U.
  \]

- \( A \) is contained in \( B \) (or \( A \) is included in \( B \)), \( (A \subseteq B) \), when
  \[
  m_A(x) \leq m_B(x) \quad \forall x \in U.
  \]

- \( A \) is the complement or negation of \( B \) \( (A = B^c) \) when
  \[
  m_A(x) = 1 - m_B(x) \quad \forall x \in U.
  \]

In addition, Ruspini[Rus 70] defined “fuzzy partition” as follows:

**Fuzzy Partition:** A collection of fuzzy sets \( A_1, \ldots, A_k \) defined on \( U \) is said to form a fuzzy partition of \( U \) when

\[
\sum_{i=1}^{k} m_{A_i}(x) = 1 \quad \forall x \in U.
\]

We will see later that there are strong reasons for not insisting on these relations.

2.2.4 Issues of Concern

Consider a fuzzy set \( A \) with \( m_A(x) \leq 0.5 \), for all \( x \). Its complement will have membership function \( m_{A^c}(x) = 1 - m_A(x) \geq 0.5 \), for all \( x \). According to the above definition of inclusion, we will have the disturbing conclusion that \( A \) is a subset of its own

\(^1\)Note that we are not talking about “fuzzy relation”, which is a fuzzy subset of the cartesian product of two sets.
complement, or "If \( A \) then not \( A \)" (even though \( A \) is not an empty set). That is, the law of self contradiction is violated. Violations of other commonly accepted rules are among the issues raised against conventional fuzzy set theory:

- Fuzzy sets do not satisfy basic rules in set theory and logic.

Watanabe[Wat 78] pointed out that fuzzy sets with \( \text{max/min} \) for union/intersection violate the following:

- The law of excluded middle, \( A \cup A^c = U \). \hspace{1cm} (i)

- The law of contradiction, \( A \cap A^c = \emptyset \). \hspace{1cm} (ii)

- The law of self contradiction, \( A \subset A^c \rightarrow A = \emptyset \). \hspace{1cm} (iii)

To see (i) and (ii), note that when \( \text{max/min} \) are the union/intersection operations,

\[
m_{A \cup A^c}(x) = \text{max}[m_A(x), 1 - m_A(x)] \neq 1
\]

and

\[
m_{A \cap A^c}(x) = \text{min}[m_A(x), 1 - m_A(x)] \neq 0
\]

unless \( A \) is a crisp set.

- Basic operations in fuzzy set theory such as fuzzy union/intersection are often chosen in an ad hoc way. In addition to the \( \text{max/min} \) operations proposed in
[Zad 65], many alternative operations are proposed due to the limited success of max/min, and there are essentially no guidelines in choosing among them. This presents difficulties especially for applications such as fuzzy pattern recognition where fuzzy union/intersection play a major role.

Both these issues are related to fuzzy union/intersection, and are further related to the basic assumption of **truth functionality**[Gai 78][Bie 92] in fuzzy set theory. **Truth functionality** means the truth value of a composite is completely determined by the truth values of its components. The individual membership functions are all that we use to specify relations between multiple fuzzy sets, and to combine fuzzy sets using unions and intersections. A result is that the same fuzzy union/intersection operations will be used to combine any two fuzzy sets. This leads to the violation of the basic rules in set theory mentioned above.

Bellman and Giertz[BeG 73] showed that, with the assumption of truth functionality, max/min are the only choices of fuzzy union/intersection operations that satisfy most basic rules in set theory (but not the law of excluded middle, for example).

Gaines[Gai 78][Gai 84] showed that the max/min fuzzy set theory\(^2\) and probability both satisfy many common rules. For example:

**Idempotency:** \( A \cap A = A \).

**Distributivity:** \((A \cup B) \cap C = (A \cap C) \cup (B \cap C)\).

Requiring in addition that the law of excluded middle be satisfied leads to probability, and requiring truth functionality leads to max/min fuzzy set theory [Gai 78]. It is possible to have truth functionality and satisfy the law of excluded middle by changing the definitions for complement and the max/min operations, but idempotency and distributivity will be violated. (See Chapter 2 in [FoK 88]).

Watanabe[Wat 78] believed that the violation of any basic laws of logic should be viewed as a weakness of fuzzy logic. Lindley[Lin 87] viewed probability as the only satisfactory tool in modeling uncertainty of any nature, including randomness and fuzziness. Bier[Bie 92] pointed out that truth functionality is an assumption rather

\(^2\)Max/min fuzzy set theory refers to the original theory according to Zadeh[Zad 65] where fuzzy union/intersection operations are max/min.
than a required property for fuzzy set theory, and it may not be suitable for linguistic reasoning.

Hisdall[His 94] and Cheeseman[Che 86] both used conditional probabilities in place of membership functions. Hisdall discussed several probabilistic experiments for soliciting membership function values, and various sources of fuzziness (e.g. measurement error, insufficient information, different opinions due to different people).

This thesis is motivated by the belief that the mathematical structure of probability can contribute to the interpretations of fuzzy sets, and provide insights into fuzzy systems. In comparison to previous works [His 86] [Che 86], our approach is different in that we start by clarifying the idea of an enlarged sample space. A clear focus on this enlarged sample space is essential to understanding how probability models can be used to interpret fuzzy sets.
Chapter 3

Probabilistic Interpretations of Fuzzy Sets

3.1 Introduction

Topics covered in this chapter are as follows.

How to model fuzzy sets probabilistically.

In 1914, before fuzzy set theory was created, Borel gave a probabilistic solution to the problem of "How many grains constitute a heap of wheat?", which involves the fuzzy concept "heap" [Bor 14]. Borel's idea of polling a population is introduced in section 3.2.

In section 3.3 the idea of constructing fuzzy sets with random intervals, suggested by Mitter [Mit 93], is discussed. Fuzzy sets with triangular shapes and trapezoidal shapes can be modeled this way very naturally. Dithered quantization is an example where a purely random phenomena now has a fuzzy interpretation due to the correspondence between fuzzy sets and random intervals.

In 3.4 we study the common structure of all probabilistic models used to describe fuzzy sets. The idea of the enlarged sample space, the foundation of probabilistic

\[1\text{This is also known as the "random set" idea and has been discussed independently elsewhere [DuP 94].} \]
models, is introduced. In any probabilistic model for fuzzy sets, the key is to **enlarge** the universe of discourse on which fuzzy sets are defined, so that fuzzy sets become ordinary events (crisp sets) in the enlarged sample space. The sample space can be viewed as the Cartesian product of the observation space and a description space to be defined later. The membership function of fuzzy set $A$ corresponds to the conditional probability $m_A(x) = P(A|x)$.

**The advantage of viewing fuzzy sets probabilistically.**

Probability has the mathematical structure to deal with various relations between fuzzy sets, and to select union/intersection operations accordingly. Fuzzy sets with probabilistic interpretations can satisfy all basic laws of logic, including the law of excluded middle. Within the framework of probability, various assumptions on the relations between fuzzy sets (events) can be made to simplify the computation.

In section 3.5 we introduce basic relations between fuzzy sets, and how they affect the operations for fuzzy union/intersection. Some typical relations between fuzzy sets in fuzzy applications are discussed.

Modeling fuzzy sets probabilistically suggests another way to describe the association between a fuzzy set and its elements. The pseudomembership function of a fuzzy set $A$ is $pm_A(x) = P(x|A)$, it measures the plausibility of $x$ appearing under the hypothesis $A$. The idea of a pseudomembership function is discussed in section 3.6. Note that the distinction between membership functions and pseudomembership functions exists only in the probabilistic framework, not in conventional fuzzy set theory.

In section 3.7, some common and useful relations for fuzzy sets describing different variables are summarized.

**Comparison with Zadeh's Results.**

In [Zad 68], Zadeh viewed fuzzy sets as "fuzzy events" (a new type of event) in a sample space, and discussed properties such as the probability of a fuzzy event. Differences between our results and Zadeh's results are discussed in section 3.8.
3.2 Borel’s Idea: Polling a Population

3.2.1 Describing a Heap of Wheat

Consider the question: “How many grains constitute a heap of wheat?”. This question has long been of interest to philosophers, as it is not easy to come up with a clear answer. The concept of a heap—defined over the numbers of grains—is clearly fuzzy. It is simply impossible to find a number of grains \( g \) such that \( g \) grains constitute a heap, but \( g - 1 \) grains are not a heap. Intuitively the degree of heapness—as a function of the number of grains \( x \)—should grow smoothly as \( x \) increases.

How do we find this function, the degree of heapness? Borel suggested taking a poll from a large population, where people have a reasonable understanding of the concept “heap”. Each person in the population is asked the question: “Do \( x \) grains constitute a heap?”. The answer must be either “yes” or “no”. The fraction of people that answer “yes” can be used as the degree of heapness of \( x \) grains. The degree of heapness will thus grow from zero towards one as the number of grains increases[Bot 14].

3.2.2 The Probabilistic Interpretation of Borel’s Idea

Based on Borel’s idea, we will now show how one can always interpret fuzzy sets in a probabilistic way.

Consider a probabilistic experiment where we randomly select, with equal probability, a person from the population to answer the question “Do \( x \) grains constitute a heap?”. The probability that the answer is “yes” will be the same as the fraction of the whole population that answer “yes”, or the degree of heapness. The degree of heapness will correspond to the probability that the randomly chosen person agrees that \( x \) grains form a heap.

To make the mathematical model complete, let’s view the number of grains as a random variable \( X \). The degree of heapness thus corresponds to the conditional probability that, given \( X = x \), the randomly chosen person agrees that there is a
heap.

The sample space of this probabilistic model could be defined as the set of pairs 
\{(x, i) | 1 \leq i \leq p, x \in \mathbb{Z}^+\} where \(x\) is the number of grains, \(i\) is the \(i\)th person in 
the population, \(p\) is the population size, and \(\mathbb{Z}^+\) is the set of positive integers. The 
sample point \((x, i)\) corresponds to the event: there are \(x\) grains, and the \(i\)th person 
is selected to answer the question “Do \(x\) grains form a heap?”. Relating to the fuzzy 
concept heap there is a function \(h(x, i)\), the yes/no answer of the \(i\)th person to the 
question “Do \(x\) grains constitute a heap?” for all \(x\) and \(i\). \(h(x, i)\) is a deterministic 
function. Let \(H\) denote the event “heap”, the event that the selected person declares 
that a selected number of grains forms a heap. \(H\) is thus the union of events \((x, i)\) 
over all \(x\) and \(i\) for which \(h(x, i) = yes\). The degree of heapness corresponds to 
the conditional probability

\[
P(H|x) = P[\bigcup_i \text{ where } h(x, i) = yes(x, i)|X = x].
\]

To see how fuzzy sets fit in, note that the fuzzy set “heap” defined over the number 
of grains \(U = \mathbb{Z}^+\) can be viewed as the event “heap” in the probabilistic context\(^2\). 
We will use \(H\) to stand for both the fuzzy set “heap” and the probabilistic event 
“heap”. A natural choice for the membership function \(m_H(x)\) will be the degree of 
heapness, i.e. the conditional probability \(P(H|x)\). In other words,

\[
m_H(x) = P(H|x).
\]

The experiment above should be regarded only as a conceptual experiment. Asking 
each member of a set of people whether each given number of grains forms a heap 
is a sensible way of constructing a membership function. But once the membership 
function \(m_H(x)\) is specified, there is no need, and no point, in actually performing the 
experiment of choosing a random individual and asking a question about a given num-
ber of grains. Thus the probabilistic interpretation is to be thought of as a conceptual

\(^2\)A more detailed explanation is given in Section 3.4
experiment rather than an actual experiment.

3.2.3 Finding the Membership Function as a Fraction of a Population

We now give a number of alternative ways of choosing a membership function $m_A(x)$. In each, the membership function $m_A(x)$ corresponds to the fraction of a population of individuals or more abstract entities that agree that the given observation $x$ has property $A$. This fraction is equal to the probability that a randomly chosen member of that population agrees that $x$ has property $A$.

- **Survey a population**: to describe one fuzzy concept $A$ probabilistically, ask each individual member of a population to give yes/no answers to “Do you think $x$ (an observation) has property $A$?”, and find the fraction of people who answer “yes”. Another way to state this question is “Do you think $x$ is $A$, or not $A$ (i.e. $A^c$)? Please choose exactly one answer.”

  The degree of truth that $x$ is $A$
  
  $= m_A(x)$
  
  $= \text{the fraction of people that choose } A$
  
  $= P(A|x)$.

- **Ask an Expert**: The population can be thought of as existing only in an expert’s mind. In that case, to find the membership function of a fuzzy set $A$, we ask the expert to specify $m_A(x)$, $\forall x$, which corresponds to the fraction of the population in his mind that think $x$ has the property $A$.

- **Ask Several Experts**: We might ask several experts for a membership function and take a weighted average as the final membership function. The opinions from some experts can be weighted more. The population is modeled as being broken down into smaller groups, one in each expert’s mind. The answer of each expert is weighted proportional to the size of his group.
For example, suppose the population size is modeled as \( p = n_1 + n_2 + \ldots + n_e \), there are \( e \) experts, and the \( i \)th expert is modeled as having population size \( n_i \). To find the membership function of fuzzy set \( A \) given observation \( x \), expert \( i \) gives \( m_i^A(x) = r_i/n_i \), the percentage of his population that think \( x \) is \( A \), \( r_i \) is the number of votes in his population that chooses \( A \). The membership function for \( A \) will be a weighted average of the membership function obtained from each expert.

\[
m_A(x) = \frac{\text{size of the population that choose } A}{\text{size of the population}} \\
\quad = \frac{\sum_{i=1}^{e} r_i}{p} \\
\quad = \frac{\sum_{i=1}^{e} n_i r_i}{p \sum_{i=1}^{e} n_i} \\
\quad = \frac{\sum_{i=1}^{e} n_i m_i^A(x)}{p}
\]

Borel's idea provides a standard way to view fuzzy concepts probabilistically. The fuzziness of whether or not \( x \) grains constitute a heap is described by the randomness in the answers, assuming only "yes/no" answers are allowed. If there is no randomness (the whole population agree on the same answer), there is no fuzziness. The maximum randomness (half of the answers are "yes") corresponds to maximum fuzziness \( m_H(x) = .5 \).

### 3.3 Mitter's Idea: Random Intervals

Mitter [Mit 93] proposed another way to construct fuzzy sets probabilistically. The idea is to associate fuzzy sets with random intervals. \(^3\)

Consider a fuzzy set \( A \) defined on the real line, i.e. the observation space is \( R \) (\( R \) is the "universe of discourse" in conventional fuzzy set theory). Let \( V, W \) be random variables associated with \( A \). \( f_{V,W}^A \) is the joint probability density of \( V \) and \( W \).

\(^3\)The random interval idea is also known as the random set idea and has been discussed independently elsewhere [DuP 94].
The membership function \( m_A(x) \) can be defined in terms of \( V \) and \( W \) as follows:

\[
    m_A(x) = \text{Prob}(x \in [V, W]) = \text{Prob}(V \leq x \leq W) = \int_{v \leq x, w \geq x} f_{V,W}^A(v, w) dv \, dw
\]  

(3.1)  

(3.2)  

(3.3)

Note that \( x \) is an arbitrary constant here, and the probability refers to the random variables \( V \) and \( W \). That is, the membership function \( m_A(x) \) is the probability that the random interval \([V, W]\) contains \( x \). Each fuzzy set is characterized by the joint density of the lower and upper limits of the random interval. The membership function for fuzzy set \( A \) is specified by \( f_{V,W}^A \). This can be easily generalized to higher dimensional fuzzy sets.

To make the probabilistic model complete, we can view \( X \) as being a random variable, independent of \( V \) and \( W \).

**Special Case:** If \( V, W \) are independent of each other, and their cumulative distribution functions are \( F_V, F_W \), then

\[
    m_A(x) = P(x \in [V, W] | x) = P(V \leq x | x)P(W \geq x | x) = P(V \leq x)P(W \geq x) = F_V(x)(1 - F_W(x))
\]  

(3.4)  

(3.5)  

(3.6)  

(3.7)

By varying \( F_V, F_W \), a large variety of fuzzy sets can be created, including any fuzzy set that increases up to some \( x_0 \), and then decreases. Some particularly popular piecewise linear fuzzy sets occurs when \( V \) is uniformly distributed in \([v_{\min}, v_{\max}]\), and \( W \) is uniformly distributed in \([w_{\min}, w_{\max}]\).

1. **Triangular Fuzzy Sets:** \( v_{\max} = w_{\min} \).

See Fig. 3.1. \( V \) is uniformly distributed in \([-1, 0]\), \( W \) is uniformly distributed in \([0, 1]\), and the triangular fuzzy set peaks at \( v_{\max} = w_{\min} = 0 \).
2. **Trapezoidal Fuzzy Sets**: \( v_{\text{max}} < w_{\text{min}} \).

   See Fig. 3.2 (a). \( V \) is uniform in \([-2, -1]\), \( W \) is uniform in \([1, 2] \). The fuzzy membership function is equal to one in \([v_{\text{max}}, w_{\text{min}}] = [-1, 1] \).

3. **Boundary Trapezoidal Fuzzy Sets**.

   See Fig. 3.2 (b). \( v_{\text{min}} = v_{\text{max}} = -\infty \) (\( V = -\infty \) with probability one). \( W \) is uniform in \([0, 1]\). This fuzzy set is on the lower boundary. The upper boundary fuzzy set will have \( w_{\text{min}} = w_{\text{max}} = \infty \).

   Another particularly simple situation occurs when the random interval has a fixed length \( c \), \( W = V + c \), but the location is random. \( V \) and \( W \) are completely dependent on each other in this case. When the random shift of the interval is uniformly distributed, we can also obtain popular fuzzy sets with triangular or trapezoidal membership functions.
Example: Dithered Quantization.

In uniform quantization with unit length quantization intervals, a time varying signal \( x_t \in \mathbb{R} \) is quantized into an integer \( Q(x_t) \) as follows.

\[
Q(x_t) = i \quad if \quad x_t \in [i - 1/2, i + 1/2].
\]

That is, \( x_t \) is quantized into integer \( i \) if it falls into to reduce the \( i \)th quantization interval, \([i - 1/2, i + 1/2]\). See Fig. 3.3.

![Dithered Quantization Diagram](image)

Figure 3-3: Dithered Quantization.

To reduce the perception of "blockiness" of the waveform \( Q(x_t) \) (\( Q(x_t) \) can be a quantized speech signal), a common approach is to "dither" the quantization intervals (see [GeG 92], page 169, problem 5.9) by a random amount \( N \) uniformly distributed in \([-1/2, 1/2]\). Each quantization interval will now be a random interval. Let \([V_i, W_i]\) be the \( i \)th quantization interval. \( V_i \) will be uniformly distributed in \([i - 1, i]\), and \( W_i = V_i + 1 \) is uniformly distributed in \([i, i + 1]\). All the quantization intervals are shifted simultaneously by the same random amount, so \( V_i = W_{i-1} \), for all \( i \). We can define a fuzzy set for each random quantization interval.

Probabilistic experiments used to describe fuzzy concepts, such as a heap of wheat, are often conceptual. In dithered quantization, the probabilistic experiment is actually performed. This is an example where a problem with pure randomness has a natural fuzzy set interpretation. It is natural to define fuzzy set \( A_i \) for the \( i \)th quan-
interval as the membership function $m_{A_i}(x)$. The fuzzy set $A_i$ contains “numbers quantized into $i$”, defined over $x_i \in R$. The membership function is a triangle that peaks at $i$ and goes down to zero at $i + 1, i - 1$ (see Fig. 3.3).

A fuzzy set like $A_i$ is often called a “fuzzy number”. Changing the range of $N$ produces other piecewise linear fuzzy sets. If the random shift is reduced or increased, $A_i$ becomes a “fuzzy interval” (i.e. a trapezoidal fuzzy set whose membership function will be flat in the center and go to zero at the two ends). Suppose each quantization interval has length 1, and the random shift is uniform in $[-\alpha / 2, -\alpha / 2]$. When $\alpha < 1$, the fuzzy interval will be “normal” (i.e. the maximum membership function value is one). The physical meaning is: some numbers will always be quantized into the same interval. If $\alpha > 1$ the maximum membership function value will be less than 1 (all numbers can be quantized into more than one interval). In any case the membership functions always add up to one, $\sum_{i=-\infty}^{\infty} m_{A_i}(x) = 1$. The quantization intervals always partition the real line, regardless of the amount of the random shift. The idea of fuzzy partition will be further discussed in section 3.5.1.

3.4 The Probabilistic Model for Fuzzy Sets

Two conceptual probabilistic models for modeling fuzzy sets (surveying a population, random intervals) have already been introduced. We next study the common features of all these probabilistic models, starting from the idea of the enlarged sample space.

3.4.1 Modeling Fuzzy Sets as Events in the Enlarged Sample Space

The sample space $S$ of a probabilistic experiment contains a collection of possible outcomes called sample points. An event in $S$ is a set of sample points, and thus an event in $S$ is a subset of $S$. The singleton events, i.e. the events consisting of single points, are mutually exclusive, collectively exhaustive, and finest grain[Dra 67]. In what follows, we often refer to these singleton events simply as sample points. A
probability space is a triplet \((S, F, P)\), where \(S\) is the sample space, \(F\) is the \(\sigma\)-field of Borel sets in \(S\) (each element in \(F\) is an event), and \(P\) is a probability measure over \(S\). Given a sample space and the probability measure, we can find the probability of any event in \(S\) (i.e. any element in \(F\)).

**The Observation Space \(U\)**

Consider defining fuzzy sets on a collection of elements \(U\). \(U\) is usually called the universe of discourse. Here, however, we refer to \(U\) as the observable universe of discourse, or the observation space. The word observable indicates that \(U\) contains items whose values (or features) can be measured or observed. Often \(U\) is a collection of numerical values or vectors, but \(U\) can contain people, characters, or other items. For example, consider \(U = \{1, 2, 3, 4\}\) (Fig. 3.4). In general, \(U\) can be multidimensional when there are multiple variables, and can be countably or uncountably infinite. Our results will be valid for any \(U\); the simple case \(U = \{1, 2, 3, 4\}\) is chosen for ease of illustration.

If a probability space \((U, F, P)\) is defined on \(U\) as a sample space, then an event in \(U\) will contain a crisp set of sample points (for example see Fig. 3.4 (a)). Each sample point is either in or not in any given event. Zadeh[Zad 68] defined a “fuzzy event” to be a fuzzy set in the sample space \(U\). A sample point can be in the “fuzzy event” to some degree.

We will take a different approach. To discuss fuzzy sets in \(U\), we will not take \(U\) as the sample space anymore. Our sample space \(S\) will be obtained by enlarging \(U\). Fuzzy sets will be viewed as ordinary events (crisp sets) in our enlarged sample space.

**The Enlarged Sample Space \(S\)**

The enlarged sample space \(S\) in our probabilistic model is obtained by replacing each sample point in \(U\) (the observation space) with multiple sample points. We use a patch in the sample space \(S\) to represent the event “\(X=x\)”, indicating that it contains multiple sample points (Fig. 3.4). The observation space \(U\) forms a partition of \(S\). In the observation space \(U\), the singleton event \(X = x\) can either be completely
Figure 3-4: Enlarge the observation space to obtain the sample space.

included or not included in any other event. \( X = x \) cannot be included by another event "to some degree". In the enlarged sample space \( S \), the event \( X = x \) contains multiple sample points, and thus can overlap with other events in \( S \) partially. The intersection of the event \( X = x \) with an event \( A \) indicates the degree to which \( X = x \) belongs to \( A \). In other words, any ordinary event (crisp set) in \( S \) can be viewed as a fuzzy set in \( U \), as it will include event \( X = x \) to some degree.
Fuzzy Sets in \( U \) are Crisp Sets in \( S \)

Definition: A fuzzy set \( A \) in the observation space \( U \) is a crisp set (an event) in the enlarged sample space \( S \) (Fig. 3.5(a)).

![Diagram of fuzzy sets in U and S](image)

Figure 3-5: Fuzzy sets are events (crisp sets) in the enlarged sample space.

In general,

- any event in \( S \) can be viewed as a fuzzy set in \( U \).

- \( A \) is a crisp set in \( U \) if each event \( X = x \) is either included in \( A \) or disjoint from \( A \).

- \( A \) is a strictly fuzzy set in \( U \) if one or more of the points \( X \in U \), viewed as events in \( S \), partially overlap with \( A \).

- Whether or not \( A \) is a strictly fuzzy set in \( U \), \( A \) is always a crisp set (event) in the sample space \( S \).

An Important Note

In the following, when the term fuzzy sets are used, we are talking exclusively about fuzzy sets defined in the observation space \( U \). They will also be referred to as events (crisp sets), in which case we are viewing them in the sample space \( S \).
3.4.2 General Structure of the Enlarged Sample Space

Our next question is: how do we actually augment \( X = x \), a point in \( U \), into multiple sample points in \( S \)? How do we introduce the additional randomness into the event \( X = x \)?

The answer depends on the conceptual probabilistic experiment. With Borel’s method each event \( X = x \) is divided into \( m \) sample points, \( \{(z, i)|i = 1, \ldots, m\} \), where \( m \) is the population size. Event \((x, i)\) happens when the \( i \)th person is selected from the population to answer a question regarding observation \( X = x \) (see section 3.2). With Mitter’s method [Mit 93] \( x \) could be divided into uncountably many sample points, one for each possible random interval. Although the details of the enlarged sample space differ for different conceptual experiments, there is also a common structure. In the following we introduce the general structure of the enlarged sample space.

The Description Space \( L \)

Suppose we are interested in \( k \) fuzzy sets \( A_1, \ldots, A_k \) (each of which is an event in \( S \)). By taking all intersections of these events and their complements, we can obtain a collection of events (fuzzy sets) \( L \) that partition the sample space \( S \). We shall call the collection of these events the description space. \(^4\)

\( L \) is the Cartesian product \( \{A_1, A_k^c\} \times \{A_2, A_k^c\} \times \ldots \times \{A_k, A_k^c\} \). An element of \( L \) is the intersection of \( k \) fuzzy sets, the \( i \)th of which is either \( A_i \) or \( A_i^c \). For example, \( l_1 = A_1 A_2 \ldots A_k, l_2 = A_1 A_2 \ldots A_k^c \), and so on. Some of these sets might be empty, but the number of elements in \( L \) will be at least \( k \) (when \( A_1, \ldots, A_k \) form a partition). Also the number of elements in \( A \) is at most \( 2^k \). Each element of the description space, which is an event in \( S \) and a fuzzy set in \( U \), will be called an elementary fuzzy set.

For example, the observation space may be \( U = [0, 1] \), and the description space \( L \) may contain fuzzy sets \{small, medium, large\}. The collection of fuzzy sets used to describe a numerical value, such as \{small, medium, large\}, might be reasonably

\(^4\)We use \( L \) to denote description space since events in \( L \) often correspond to linguistic descriptions of numerical values in \( U \).
modeled as forming a description space, i.e. partitioning the enlarged sample space\textsuperscript{5}.

Note however, that fuzzy sets we are interested in do not always form a description space. For example, \{tall, thin\} defined over people does not form a description space, as a person can be both tall and thin to high degrees or neither tall nor thin. The sample space is clearly not partitioned by events "tall" and "thin". However, we can create a description space that contains the following elementary fuzzy sets: \{tall and thin, tall and not thin, not tall and thin, not tall and not thin\}.

![Diagram](image)

Observation space \( U = \{1, 2, 3, 4\} \) : solid lines
Description Space \( L = \{ AB, A^c B, A^c B^c, A B^c, B^c \} \): dotted lines
\( A, B \) : fuzzy sets of interest, \( c \) : complement
Sample space \( S = U \times L = \{(x, l) \mid x \in U, l \in L\} \)

Each patch is a sample point

Figure 3.6: Sample space as the Cartesian product of observation space and description space.

**Sample Space as the Cartesian Product of Observation Space and Description Space**

We can now view the **enlarged sample space** \( S \) as the **Cartesian product** of the **observation space** and the **description space**, i.e. \( S = U \times L = \{(x, l) ; x \in U, l \in L\} \), see Fig. 3.6.

Given a probability measure on \( S \), the **membership function** of a fuzzy set \( l \in L, m_l(x) \), is taken as the conditional probability of event \( l \) given \( x \). \textsuperscript{6}

\textsuperscript{5}Hisdal [His 94] referred to the elementary fuzzy sets that form a description space as "a collection of complete, nonredundant label sets".

\textsuperscript{6}Note that, for any crisp set \( A \) in the observation space \( U \), the membership function \( m_A(x) \) has always been the conditional probability \( P(A|x) \). \( P(A|x) = 1 \) if \( x \in A \) and \( P(A|x) = 0 \) otherwise, regardless of \( P(x) \).
If $A_1, A_2, \ldots A_k$ are fuzzy sets of interest that generate $L$, then the enlarged sample space as defined by $S = U \times L$ is “canonical” in the sense that it is the smallest sample space in which arbitrary unions, intersections, and complements of $A_1, A_2, \ldots A_k$ and events in $U$ exist.

The Prior Probability

In classical fuzzy set theory, one only defines $U$, the fuzzy sets, and the membership functions for the fuzzy sets. In our probabilistic formulation, on the other hand, we define a probability measure on $S$, which specifies in addition the membership functions on all intersections, unions, and complements of fuzzy sets, and an apriori distribution on $U$.

Something of the nature of apriori probabilities seems to be implicit in most uses of fuzzy sets, particularly where the observation space is unbounded (e.g. $U = R$ or $R^n$). For example, in fuzzy control the input rarely exceeds some known range, and that affects the choice of the fuzzy membership functions in the sense that none of the membership functions are varying outside of that range. This range and its inherent fuzziness can be visualized by an apriori probability measure.

At any rate, one can justify the added complexity of an apriori distribution by the fact that it gives one a complete mathematical framework. If one wishes to omit the apriori measure on $U$, however, one gets something akin to the information theoretic channel with input alphabet $U$ and output alphabet $L = \{A_1, A_i^j\} \times \{A_2, A_j^2\} \times \ldots \times \{A_k, A_k^k\}$.

3.4.3 Visualization of the Enlarged Sample Space

Discrete Observation Space

When $U$ and $L$ are both small we might visualize the enlarged sample space in a unit square as in Fig. 3.7 (a), (b). The enlarged universal set is a one by one square. Each rectangular patch in the square corresponds to a sample point $(x, l)$, and the area of each patch is the probability of that sample point.

In Fig. 3.7(a) the universal set is partitioned first by events $x \in U$ (observations),
Figure 3-7: Visualize the enlarged sample space: discrete observation space.
then subdivided by events $l \in L$ (elementary fuzzy sets). The unit square is divided into vertical strips, one for each $x \in U$. The width of each strip corresponds to probability $P(X = x)$, which sum to one over $x \in U$. Each strip (an event $x \in U$) is further divided horizontally into patches by $l \in L$. The height of each patch corresponds to $P(l|x)$, and these heights sum to 1 over $l \in L$.

In Fig. 3.7(b) the universal set is partitioned first by events $l \in L$ (elementary fuzzy sets), then subdivided by events $x \in U$ (observations). The width and height of a rectangular patch here corresponds to $P(l)$ and $P(x|l)$, respectively.

The area of each patch $(x, l)$ in either Fig. 3.7(a) or (b) is the probability of that sample point $P(x, l) = P(x)P(l|x) = P(l)P(x|l)$.

Continuous Observation Space

We now generalize to the case in which the observation space is continuous, e.g. $U = R$. The enlarged sample space can still be visualized in a unit square, see Fig. 3.8. In Fig. 3.8, each event $x \in U$ corresponds to a vertical line located at a distance of $P(X < x)$ from the left boundary. The line is further partitioned into segments by fuzzy sets, the length of each segment corresponds to the corresponding membership function value.

The Symmetry of Observation Space and Description Space

The two ways to visualize the sample space show the symmetry between the description space $L$ and the observation space $U$. We note the following:

- $L$ and $U$ each forms a partition of the enlarged sample space. Whereas small is a fuzzy set in $U$, 3 could also be viewed as a fuzzy set defined in $L = \{small, large\}$. Each $x \in U$ overlaps with $l \in L$ to some degree, and vice versa.

- The major difference between $U$ and $L$ is that the cardinality of $L$ is usually smaller than that of $U$, as we summarize the observations in $U$ (often numerical values) with linguistic descriptions in $L$ (small, large, etc).
In general, for any two partitions of the sample space, one can be viewed as a collection of elementary fuzzy sets defined over the other and vice versa. A special case: if $U$ can be obtained by further partitioning $L$, then each $l \in L$ is a crisp set over $U$.

We mentioned before that the enlarged sample space (i.e. the product of observation space and description space) is the smallest sample space from which the probabilities of all events of interest can be found. The sample space in Borel's experiment (Fig. 3.9), for example, contains more elements.

Defining the sample space with all events of interest so that the number of sample points is the smallest, as we did, may not always be desirable. Consider an experiment where a coin is tossed 100 times independently, and all that interests us is the total number of heads. The sample space, if defined with events of interest, consists of sample points "the total number of heads is $n", \ 0 \leq n \leq 100$. However, due to the structure of this experiment (the results for different tosses are independent and identically distributed), a better definition is to have as a sample point the sequence
\{x_1, \ldots, x_{100}\}$, where $x_i$ is 1 if the result of the $i$th toss is head, otherwise $x_i = 0$. This sample space is less compact, but it allows us to find probabilities of events more easily.

### 3.4.4 Finding Membership Functions for Multiple Fuzzy Sets

We discussed selecting the membership function for one fuzzy set by surveying a population or asking experts in Section 3.2.3. We generalize in the following to multiple fuzzy sets, in particular the elementary fuzzy sets in the description space.

Suppose the description space $L$ contains elementary fuzzy sets $l_1, l_2, \ldots, l_k$. The membership functions $m_{i_1}, m_{i_2}, \ldots, m_{i_k}$ can be found as follows.

- **Survey a Population**: Ask each member of a population to select exactly one answer to the question: "Do you think $x$ has property $l_1$, or $l_2$, \ldots, or $l_k$?", for all $x \in U$. The fraction of people choosing $l_i$ is $m_{i}(x)$, which corresponds to $P(l_i|x)$, the probability a randomly chosen person decides that $x$ has property
\[ l_i. \]
\[ \sum_{i=1}^{k} m_{l_i}(x) = 1, \quad \forall x. \]

- **Ask an Expert:** Viewing the population as existing in an expert’s mind, we can simply ask the expert to specify \( m_{l_i}(x) \) for \( i = 1, \ldots k \) and \( \forall x \in U \), such that \( \sum_{i=1}^{k} m_{l_i}(x) = 1, \forall x \).

- **Ask several Experts:** We can ask several experts to provide the membership functions and take a weighted average of them. The population of size \( p \) is viewed as being broken down among \( e \) experts into groups of sizes \( n_1, \ldots n_e \), \( \sum_{i=1}^{e} n_i = p \). From expert \( i \) we have \( m_{j_i}^i(x), j = 1, \ldots k \), and \( \sum_{j=1}^{k} m_{j_i}^i(x) = 1 \).

We then take the weighted sum. For example,

\[ m_{l_i}(x) = \sum_{i=1}^{e} \frac{n_i}{p} m_{j_i}^i(x). \]

Note that the resultant membership functions still add up to one, since for each expert, the membership functions she/he provides add up to one:
\[
\begin{align*}
\sum_{j=1}^{k} m_{j_i}(x) \\
= \sum_{j=1}^{k} \sum_{i=1}^{e} \frac{n_i}{p} m_{j_i}^i(x) \\
= \sum_{i=1}^{e} \frac{n_i}{p} \sum_{j=1}^{k} 1 \\
= 1.
\end{align*}
\]

### 3.4.5 Fuzzy Interpretations of Random Phenomena

Probability can be used to deal with both randomness that is real (e.g. noise in communications) and randomness created for modeling purpose (e.g. when modeling fuzzy set “small” probabilistically). Probability provides a unifying framework for dealing with fuzziness and randomness. Knowledge and insights in one theory can be easily translated into the other, which is beneficial to both theories.

Dithered quantization in section 3.3 is one example where a problem with real randomness has a fuzzy interpretation. In general, whenever the sample space is partitioned in two ways, the collection of events that forms a coarser partition can be
viewed as fuzzy sets defined on the events that form a finer partition. Many problems fall into this category.

For example, consider binary communication over a channel with added white gaussian noise.

- Transmitted signal: $Y = -1$ or $1$ with equal probabilities $P_Y(-1) = P_Y(1)$.

- Received signal: $X = Y + N$, $N$ is a zero mean, unit variance Gaussian random variable independent of $Y$.

This is the most basic “signal plus noise” problem. The probabilities are shown in Fig. 3.10.

Figure 3-10: Probabilities in the binary communication problem.

The events $Y = -1$ and $Y = 1$ can be viewed as fuzzy sets defined over $X \in R$. The sample space $S$ is $R \times \{Y = -1, Y = 1\}$. $S$ is partitioned both by the observation space $U = R$ and the description space $L = \{Y = -1, Y = 1\}$. $L$ is a coarser partition of $S$, and $U = R$ forms a finer partition. The a posteriori probabilities are a natural choice for the membership functions. $m_{Y=-1}(x) = P_{Y|X}(-1|x)$, $m_{Y=1}(x) = P_{Y|X}(1|x)$. Given the received signal $x$, the maximum a posteriori probability decision rule (MAP) is to select $y$ that maximizes the posterior probability $P_{Y|X}(y|x)$, which
corresponds to finding \( y \) that maximizes the membership function value. This agrees with fuzzy pattern recognition. In fact, the probabilistic interpretations suggest that fuzzy pattern recognition can be performed as Bayesian detection. Chapter 5 contains more detail.

Note that the probability density \( P_{X|Y}(x|1) \) given in Fig. 3.10, which peaks at \( X = 1 \), gives a more intuitively appealing description of the association between \( x \) and the fuzzy set \( Y = 1 \) than the aposteriori probability \( P_{Y|X}(1|x) \). This has to do with the concept of pseudomembership function, which will be introduced in section 3.6.

### 3.4.6 Viewing Fuzzy Sets as Events (Sets), Not Functions

![Figure 3-11: Fuzzy sets viewed as sets in the enlarged space.](image)

Fuzzy sets are conventionally viewed as functions. When \( U \) is finite, fuzzy set \( A \) is viewed as \( m_A(x) \), which is a bar graph (see Fig. 3.11(a)). We suggest viewing fuzzy set \( A \) as an event (crisp set) in the enlarged sample space, as shown in Fig. 3.11(b). The association of the fuzzy set and its elements can be described in two ways by either the membership function \( m_A(x) \), which corresponds to \( P(A|x) \), or the pseudomembership function \( P(x|A) \) (see section 3.6). The probability of observing \( X = x \), \( P_X(x) \), is also included in the picture as the width of each vertical strip.
This new way to view fuzzy sets (which is also a new way to view crisp sets) is even more useful with multiple fuzzy sets, as their relations become apparent (see the next Section).

3.5 Relations between Fuzzy Sets: Our View

A major advantage of the probabilistic interpretation is the ability to deal with relations between fuzzy sets. The relations determine how fuzzy sets are combined through fuzzy union and intersection. By “relations” we are referring to properties between fuzzy sets that allow us to specify their union and intersection from the individual fuzzy sets, such as “identical”, “disjoint”, “independent”, “inclusion” (containment). We are not dealing with “fuzzy relation”, which is defined as a fuzzy subset of the cartesian product of two crisp sets.

In conventional fuzzy set theory, the relation between two fuzzy sets $A, B$ is determined solely by their individual membership functions, $m_A, m_B$. For example, $A, B$ are “identical” if $m_A(x) = m_B(x), \forall x$. In the probabilistic view, this is not necessarily valid. The relation between two fuzzy sets in general cannot be specified by the individual membership functions.

In the most general sense, “knowing the relation between $A, B$” means being able to specify $m_{A\cap B}, m_{A\cup B}, m_{A\cap B^c}, m_{A\cup B^c}$. If we already know $m_A$ and $m_B$, then all we need to know in addition to compute the things mentioned above is either $m_{A\cap B}$ or $m_{A\cup B}$. In general we cannot find $m_{A\cap B}(x) = P(A \cap B|x)$ from $m_A(x) = P(A|x)$ and $m_B(x) = P(B|x)$, unless $A, B$ are known to be related in some special way.

In the following we discuss some of these special relations. We will point out the differences between our definitions and those in fuzzy set theory.

3.5.1 The Basic Relations

What's new about relations between fuzzy sets compared to relations between crisp sets are the additional varieties of conditional relations. Fuzzy sets $A, B$ can be viewed as the union of component events(crisp sets) conditional on the observation
Figure 3-12: Two disjoint fuzzy sets.

$x$, $A = \bigcup_x (x \cap A)$ and $B = \bigcup_x (x \cap B)$. It is the relation between component events $(x \cap A)$, $(x \cap B)$, or the relation between $A$, $B$ conditional on $x$, that determines how $m_{A \cap B}(x)$ can be obtained from $m_A(x)$ and $m_B(x)$. If the conditional relations are different for different $x$, the operation for intersection can be different for different $x$. The simpler case is when all component events $(x \cap A)$, $(x \cap B)$ are related in the same way. In this case the union/intersection operations will be the same for all $x$, and we can talk about the overall (unconditional) relations between the fuzzy sets $A$, $B$. The same conditional relation does not always imply the same unconditional relation, see the examples below.

**f Disjoint:** $A \cap B = \emptyset$.

See Fig. 3.12. $A$, $B$ are disjoint if and only if $(x \cap A)$ and $(x \cap B)$ are disjoint, for all $x$. That is,

**Conditionally Disjoint $\Leftrightarrow$ Disjoint.**

A fuzzy set and its complement are always disjoint, and their union is always the enlarged space $S$. Fuzzy sets always satisfy the law of excluded middle in our interpretation. In general:
\[ A \cap B = \emptyset \Leftrightarrow (x \cap A) \cap (x \cap B) = \emptyset \quad \forall x. \]

The union/intersection operations will be:

\[ m_{A \cup B}(x) = m_A(x) + m_B(x); \quad m_{A \cap B}(x) = 0. \]

**Conditional Inclusion:**

For each \( x \), either \((x \cap A) \subset (x \cap B)\) or \((x \cap B) \subset (x \cup A)\). As a special case, if \((x \cap A) \subset (x \cap B)\) for all \( x \), then \( A \subset B \) (**Inclusion**), and \( A \subset B \) implies \((x \cap A) \subset (x \cap B)\) for all \( x \).

![Diagram](image)

**Figure 3-13:** Inclusion (a) and conditional inclusion (b).

See Fig. 3.13 (a) for an example of \( A \subset B \). In general:

\[ A \subset B \Leftrightarrow (x \cap A) \subset (x \cap B) \quad \forall x; \]

\[ \Rightarrow m_A(x) \leq m_B(x), \quad \forall x. \]

The union/intersection operations are:

\[ m_{A \cap B}(x) = \min(m_A(x), m_B(x)) \]

\[ m_{A \cup B}(x) = \max(m_A(x), m_B(x)) \]
See Fig. 3.13 (b) for an example of conditional inclusion. Note that:

\[
\text{Conditional Inclusion} \iff \text{Inclusion},
\]

but not the other way round. As long as conditional inclusion is satisfied, the union/intersection operations will be the same, max/min, for all \( x \). In conventional max/min fuzzy set theory, all fuzzy sets are effectively restricted to be related as conditional inclusion.

In conventional fuzzy set theory, \( m_A(x) \leq m_B(x), \forall x \iff A \subset B \). We agree that \( m_A(x) \leq m_B(x) \forall x \iff A \subset B \), but not the other way round. See Fig. 3.14 for an example where \( m_A(x) \leq m_B(x) \) but \( A \) is not a subset of \( B \) (actually, \( A \) is the complement of \( B \)).

![Diagram showing conditional inclusion](image)

\[\begin{align*}
A &= B^c \\
m_A(x) &< m_B(x), \forall x \\
\text{but } A &\not\subset B
\end{align*}\]

**Figure 3.14**: \( m_A(x) \leq m_B(x) \forall x \), but \( A \) is not a subset of \( B \).

**Independence:**

As we shall see, conditional independence of \( A, B \) given \( x \), for all \( x \) does not necessarily imply \( A \) and \( B \) are independent unconditionally. Unconditional independence of \( A, B \) doesn't guarantee conditional independence of \( A, B \) given \( x \), either. See Fig. 3.15, Fig. 3.16.
Figure 3-15: Two fuzzy sets with the same membership functions that are independent conditional on the observation. $P(AB|x) = P(A|x)P(B|x)$.

1. **Conditional Independence Given x**: $A$ and $B$ are conditionally independent given $x$ (Fig. 3.15) if and only if

$$P(A \cap B|x) = P(A|x)P(B|x), \forall x.$$ 

The union/intersection operations are:

$$m_{A \cap B}(x) = m_A(x)m_B(x), \forall x$$

$$m_{A \cup B}(x) = m_A(x) + m_B(x) - m_A(x)m_B(x), \forall x.$$  

2. **Overall Independence**: $A$ and $B$ are independent unconditionally (Fig. 3.16) if and only if $P(A \cap B) = P(A)P(B)$

$$\Leftrightarrow \sum_x m_{A \cap B}(x)P(x) = \sum_x (m_A(x)P(x))\sum_x (m_B(x)P(x)).$$

Zadeh defined $AB$, the "product", and $A \oplus B$, the "sum", as follows:

$m_{AB}(x) = m_A(x)m_B(x)$; and $m_{AB}(x) = m_A(x) + m_B(x) - m_A(x)m_B(x)$. In our
Figure 3.16: Two fuzzy sets with the same membership functions that are independent unconditionally. $P(AB) = P(A)P(B)$.

view, when $A$ and $B$ are conditionally independent given $x$, Zadeh’s $AB$ corresponds to our $A \cap B$ and his $A \oplus B$ corresponds to our $A \cup B$.

Identical:

In conventional fuzzy set theory, two fuzzy sets $A$, $B$ with the same membership functions are considered “identical”, i.e.

$$m_A(x) = m_B(x), \forall x \iff A = B.$$ 

With the probabilistic view,

$$A = B \Rightarrow m_A(x) = m_B(x), \forall x,$$

but not the other way round.

In Fig. 3.15 and Fig. 3.16 fuzzy sets $A$, $B$ have the same membership functions.
Figure 3.17: Three fuzzy sets (events) that form a fuzzy partition (partition the sample space).

Viewed as events, $A$, $B$ not only are not identical, but they are independent in Fig. 3.16, and conditionally independent in Fig. 3.15.

**Partition:**

$A$ and $B$ partition the sample space if they are mutually exclusive ($A \cap B = \emptyset$, $A$ and $B$ are disjoint) and collectively exhaustive ($A \cup B = S$). This happens if and only if $A$ is the complement of $B$. In our view, what “fuzzy partition” means is that the fuzzy sets partition the enlarged sample space. That is, $A \cup B = S$, and $A \cap B = \emptyset$.

In this case,

$$m_{A \cap B}(x) = 0$$

$$m_{A \cup B}(x) = m_A(x) + m_B(x) = 1, \ \forall x \in U.$$  

In general, a collection of fuzzy sets (events) $\{A_1, \ldots, A_k\}$ form a fuzzy partition if they partition the enlarged sample space. In other words, $\{A_1, \ldots, A_k\}$ are elementary fuzzy sets that form a description space. See Fig. 3.17.

According to Ruspini[Rus 70]'s original definition, fuzzy sets $A_1, \ldots, A_k$ are said to form a fuzzy partition if

$$\sum_{i=1}^{k} m_{A_i}(x) = 1, \ \forall x.$$  

In our view, “sum of the membership functions equals one everywhere” is a nec-
necessary but not sufficient condition for fuzzy sets to form a fuzzy partition.

For example, do the fuzzy sets in Fig. 3.18 form a fuzzy partition (i.e. partition the enlarged sample space), as the membership functions add up to one everywhere in $U = [0, 10]$? They may or may not. Consider the two cases:

- **Case 1:** Using the random interval idea, let $V$ be uniformly distributed in $[0, 5]$ and $W$ be uniformly distributed in $[5, 10]$. Define fuzzy membership functions as follows.

1. $m_{A_1}(x) = P(x < V|\omega)$,
2. $m_{A_2}(x) = P(V < x < W|\omega)$,
3. $m_{A_3}(x) = P(x > W|\omega)$.

$A_1, A_2, A_3$ form a partition in this case. Note that the same $V$ and $W$ are used to define all three fuzzy sets.

- **Case 2:**

1. $V_1 = -\infty$ with probability one, $W_1$ is uniform on $[0, 5]$,
2. $V_2$ is uniform on $[0, 5]$, $W_2$ is uniform on $[5, 10]$,
3. $V_3$ is uniform on $[5, 10]$, $W_3 = \infty$ with probability one.

All the $V_i, W_i$ are mutually independent, $i = 1, 2, 3$.

$$m_{A_i}(x) = P(x \in [V_i, W_i]|X = x), \quad i = 1, 2, 3.$$ 

$A_1, A_2, A_3$ do not form a partition in this case. In fact they are conditionally independent given $x$, which means they cannot be disjoint and thus cannot form a partition.

"Small" and "Large": Partition or Not?

Fuzzy sets "small" and "large" may or may not form a partition. See Fig. 3.19. The observation space is $U = [0, 1]$. In Fig. 3.19 (a), "small" and "large" are complements.
Figure 3-18: Three fuzzy sets whose membership functions add up to one.

Figure 3-19: Fuzzy Sets "Small" and "Large" on $U = [0, 1]$.  

of each other. In Fig. 3.19 (b), "small" and "large" do not form a partition, even though the membership functions might still add up to one.

Some people have argued that fuzzy sets do not need to satisfy the law of excluded middle and the law of contradiction since "a rose can be both red and not red to some degree" [BeG 73][Kos 92]. In our view a fuzzy set and its complement should always be modeled as being disjoint. When people say "$x$ is both small and not small" the intended meaning is often "the degree to which $x$ is small and the degree to which $x$ is not small are both greater than zero", or "$x$ is both small (to some degree) and not small (to some degree)". This is different from saying that the degree to which $x$ is in the fuzzy set "both small and not small" is greater than zero. Similarly, "$x$ is neither small nor not small" can simply mean that "the degree to which $x$ is small and the degree to which $x$ is not small are both less than one".

59
"Small" and "Very Small", Inclusion or Disjoint?

The relation between fuzzy sets "very small" and "small" also depends on the situation. Intuitively, "very small" should imply "small". When the membership function of "very small" is obtained by squaring the membership function of "small", this is the implicit idea. very small ⊆ small.

Alternatively, in many applications we have a collection of fuzzy sets such as \{very small, small, medium, large, very large\} that form a description space, with membership functions that add up to one. That is, these fuzzy sets form a partition (fuzzy partition in conventional terms). This means that they are disjoint, including "very small" and "small". In this case "small" should be interpreted as "small but not very small". "Very small" can thus be disjoint from "small but not very small". The description space really consists of \{very small, small (but not very small), medium, large (but not very large), very large\}.

Special Case: Relations Between Crisp Sets

Consider two crisp sets \(A, B\) in \(U\). By viewing them as events in the enlarged sample space \(S = U \times L\), we can obtain insight into why there are some union/intersection operations that always work for crisp sets, but lead to violation of basic rules for fuzzy sets.

For crisp sets, the membership functions \(m_A(x) = P(A|x)\) and \(m_B(x) = P(B|x)\) can only take on values 1 or 0. The component events \((x \cap A)\) and \((x \cap B)\) must either be \(x\) or \(\emptyset\). \((x \cap A)\) and \((x \cap B)\) must be either identical, or one of them is \(x\), and the other is \(\emptyset\). Regardless of how \(A, B\) are related overall (disjoint, independent, inclusion), their conditional relations will always satisfy both conditional independence and conditional inclusion. This explains why \(\max/\min\) and \(\text{sum minus product/product}\) always work for crisp sets. These operations do not always appropriate when \(A, B\) are fuzzy, since the conditional relations do not necessarily hold anymore.
3.5.2 The Choice of Fuzzy Union/Intersection Operations

1. Unknown Relations.

In our view, the max/min operations for union/intersection are appropriate only when fuzzy sets are related as conditional inclusion. However, they are reasonable choices when relations between fuzzy sets are unknown, as they provide lower/upper bounds for the actual values.

For all \( x \),

\[
0 \leq m_{A \cap B}(x) \leq \min(m_A(x), m_B(x)) \leq \max(m_A(x), m_B(x)) \leq m_{A \cup B}(x) \leq 1.
\]

2. Simple relations.

With the probability model, we have a choice of simple union/intersection operations: sum/zero if the fuzzy sets are disjoint, sum-minus-product/product if the fuzzy sets are conditionally independent, and max/min if the fuzzy sets satisfy conditional inclusion. We can often reasonably assume that fuzzy sets in our problem are related in some particular ways to simplify computation.

3. More complicated relations.

When the relations are more complicated, fuzzy union/intersection operations may no longer be as simple as the ones mentioned above. With the probabilistic model, however, we have the ability to describe completely arbitrary relations. In the next section we will look at the case in which there are more than one variables being observed, and each is described by a collection of fuzzy sets.

3.5.3 Conditional Independence of Fuzzy Sets Describing Different Observations

Suppose fuzzy sets \( A_1, \ldots, A_n \) describe numerical variable \( x \), and fuzzy sets \( B_1, \ldots, B_m \) describe numerical variable \( y \). Even though \( x \) and \( y \) can be highly dependent and \( A_i \) and \( B_j \) will also be dependent, sometimes it is reasonable to assume that given \( x \), \( A_i \) will be independent of \( y \) and any fuzzy set \( B_j \). Similarly, \( B_j \) given \( y \) will be indepen-
dent of $x$ and $A_i$. In this case, $m_{A_i \cap B_j}(x, y) = P(A_i \cap B_j|x, y) = P(A_i|x)P(B_j|y) = m_{A_i}(x)m_{B_j}(y)$. Essentially, this is the usual meaning of $A_1, \ldots A_n$ describing $x$ and $B_1, \ldots B_m$ describing $y$.

Consider a fuzzy controller with two input parameters, the temperature error ($x$, current temperature minus the target temperature) and the temperature derivative ($y = x'$). The observation space $U$ contains all possible pairs $(x, y)$. The controller might have fuzzy if-then rules such as “If the temperature error is positive large and the temperature derivative is negative small then output (something)”.

Let $A$ be any fuzzy set describing the temperature error $x$, and $B$ be any fuzzy set describing the temperature derivative $y$. We assume that given $x$, $A$ will be independent of $y$ and $B$. Similarly, given $y$, $B$ is conditionally independent of $x$ and $A$.

\[
P(A|x, y) = P(A|x) \\
P(B|y, x, A) = P(B|y) \\
P(A \cap B|x, y) = P(A|x, y)P(B|y, x, A) \\
\quad \quad \quad = P(A|x)P(B|y). \\
m_{A \cap B}(x, y) = m_A(x)m_B(y)
\]

When will $A$ and $B$, as defined above, be unconditionally independent? Intuition suggests that this should happen when the numerical variables they describe, $x$ and $y$, are themselves independent. This is indeed the case, as shown below.

\[
P(A \cap B) = \sum_{x,y} P(A \cap B|x, y)P(x, y) \\
\quad = \sum_{x,y} P(A|x)P(B|y)P(x)P(y) \\
\quad = \sum_{x,y} P(A|x)P(x)P(B|y)P(y)
\]
\[
= \sum_x P(A|x)P(x) \sum_y P(B|y)P(y) \\
= P(A)P(B)
\]

### 3.6 The Pseudomembership Function

#### 3.6.1 What is a Pseudomembership Function and When Should It be Used

Sometimes the "membership function" in conventional fuzzy set theory corresponds more naturally to \(P(x|A)\) (often multiplied by a constant to make the maximum value one) than to \(P(A|x)\). To make the distinction clear, we define the pseudomembership function of fuzzy set \(A\), \(pm_A(x)\), as follows.

\[
pm_A(x) = P(x|A).
\]

The pseudomembership function is a probability or probability density over the observations \(x\).

\[
\sum_x pm_A(x) = 1.
\]

On the other hand, the membership function is a probability over fuzzy sets in the description space, but not over the observations.

\[
m_A(x) + m_{A^c}(x) = 1.
\]

From Bayes rule,

\[
pm_A(x)P(A) = P(x|A)P(A) = P(A|x)P(x) = m_A(x)P(x).
\]

Fuzzy sets deal with the conversion between observations (often numerical and more specific) and descriptions (often linguistic and less specific). Suppose the fuzzy set of interest is \(A\). In the probabilistic model, we can either specify the pseudomembership functions \((P(x|A), P(x|A^c))\) and the prior probabilities of fuzzy sets
(P(A), P(A^c)), or the membership functions (m(A|x), m(A^c|x)) and the probability (density) of the observations (P(x)). When is it more natural to define pseudomembership functions? This depends on whether the fuzzy set plays any role in how the observation is generated.

3.6.2 Quantization-type Fuzzy Sets and Detection-type Fuzzy Sets

Quantization-type fuzzy sets: membership functions natural.

When the situation is best viewed as the occurrence of an observation followed by a fuzzy set description, fuzzy sets "describe" or "quantize" the observations, but do not play any role in their generation. The association between the fuzzy set and the observations is more naturally described by the membership function.

Example: Given an observation x (a number), the degree that x is small is m_{small}(x). Fuzzy sets in dithered quantization, and input fuzzy sets of a fuzzy controller are all quantization-type fuzzy sets.

Detection-type fuzzy sets: pseudomembership functions more natural.

When the situation is best viewed as an occurrence of a fuzzy set with an observation generated from that fuzzy set, the fuzzy set is the underlying hypothesis, as in the problem of detection. In this case, it is easier and more useful to define the pseudomembership function (the likelihood probability) than the membership function (the aposteriori probability). On the other hand, the membership function can be found using Bayes rule if the pseudomembership functions and prior probabilities of all fuzzy sets in the description space are known. Therefore, defining pseudomembership functions will also define the membership functions indirectly.

Example: Consider the binary communication problem. The transmitted signal \( Y = -1 \) or \( Y = 1 \) can be viewed as fuzzy sets over the received signal \( X = Y + N \), where \( N \) is an independent Gaussian noise. In this case the fuzzy set \( Y \) occurred
first, then an observation \( X \) is generated based on the fuzzy set. The collection of observations \( X = x \) associated with fuzzy set \( Y = y \) is more naturally described by \( pm_{Y=y}(x) = p(x|y) \), which peaks at \( x = y \). Output fuzzy sets in a fuzzy controller, and fuzzy sets in text or speech recognition are all detection-type fuzzy sets.

Given a fuzzy set \( A \), we might want to find a template or ideal element \( \mu_A \) of the fuzzy set. This "template" is often found by **defuzzification**. In our view, the pseudomembership function should be used for defuzzification. Suppose the fuzzy set is "small" and we want to find a "typical small number" by defuzzifying it. "How small is \( x \)" (\( m_{\text{small}}(x) \)) and "how often is \( x \) observed" (\( P(x) \)) should both be involved in choosing a typical small number; this can be represented by the pseudomembership function.

One way to defuzzify is to give the mean (CENTROID defuzzification):

\[
\mu_A = \sum_x pm_A(x)x = \sum_x P(x|A)x = E(X|A).
\]

Another choice is to take the "mode" of \( pm_A(x) \) (MAXIMUM defuzzification):

\[
\mu_A = \arg \max_x pm_A(x).
\]

The difference between membership functions and pseudomembership functions exists only with probabilistic interpretations, but not in conventional fuzzy set theory. Conventional fuzzy set theory can be viewed as making the assumption that \( P(x) \) is uniform \(^7\).

By looking at the membership function \( m_A(x) \) we can tell whether or not \( A \) is fuzzy (it is not if \( m_A(x) \) is either zero or one). On the other hand, the pseudomembership function \( pm_A(x) \) alone does not indicate whether \( A \) is fuzzy. For pseudomembership functions, the fact that \( A \) is fuzzy translates into: there exists \( x \) where \( pm_A(x) > 0 \) and \( pm_{A^c}(x) > 0 \).

\(^7\)The assumption that \( P(x) \) is uniform can be made if \( x \) takes on values in a finite interval. It cannot be made if \( x \) can take on any value in \( R \).
3.6.3 Conditional Independence of Observations Given Fuzzy Sets Describing Them

For detection-type fuzzy sets that are underlying hypotheses, another type of conditional independence that is dual to the situation in section 3.5.2 occurs.

Consider two observations $x, y$. Assume that $x$ is generated based on fuzzy set $A$, and $y$ is generated based on fuzzy set $B$. An example is text recognition. $x$ and $y$ could be two adjacent observed handwritten characters. $A, B$ each correspond to a conceptual character from the collection of $a$ to $z$. Each conceptual character is modeled as a fuzzy set in the pattern space, the space of all possible observations of handwritten characters. It is more natural in this case to define the pseudomembership functions $pm_A(x), pm_B(y)$.

Although $A$ and $B$ are usually highly dependent, it is reasonable to assume that, given $A$, $x$ depends solely on $A$ and nothing else, including $y$ and $B$; similarly, given $B$, $y$ does not depend on $x$ and $A$. In other words, the pseudomembership function of $A \cap B$ can be factored into the product of the pseudomembership functions.

\[
pm_{A \cap B}(x, y) = P(x, y|A, B) \\
= P(x|A, B)P(y|A, B, x) \\
= P(x|A)P(y|B) \\
= pm_A(x)pm_B(y).
\]

The purpose of character recognition is to find (detect) the most likely fuzzy sets (conceptual characters) given the observed characters. We thus need to find the membership function $P(A, B|x, y)$, which measures the likelihood of fuzzy sets given the observations, in order to make a decision. $P(A, B|x, y) = P(x, y|A, B)P(A, B)$. The "prior" probability $P(A, B)$ contains the conceptual information, the dependency between neighboring characters. The fact that the pseudomembership function (likelihood probability) can be factored makes it easy to combine component decisions
associated with each observation, and take into account their dependency through the joint prior probability at the end. See Chapter 5 for more details.

3.7 Common Relations for Fuzzy Sets Describing Multiple Variables

Suppose there are two numerical variables $x$, $y$.

$x$ is described by fuzzy sets $\{A_1, \ldots, A_n\}$,

$y$ is described by fuzzy sets $\{B_1, \ldots, B_m\}$.

In summary, some common and useful ways to relate these fuzzy sets are as follows:

1. **Partition.** $\{A_i\}$ and $\{B_j\}$ each form a description space over $x$ and $y$, respectively. That is, $\{A_i\}$ and $\{B_j\}$ each partition the corresponding enlarged sample space, or each form a "fuzzy partition". As a result,

\[
\sum_i m_{A_i}(x) = 1.
\]

\[
\sum_j m_{B_j}(y) = 1.
\]

2. **Conditional Independence.**

   (a) Quantization-type fuzzy sets: $P(A_i \cap B_j|x, y) = P(A_i|x)P(B_j|y)$.

   Fuzzy sets $A_i$, $B_j$ are conditionally independent given the numerical variables they describe. This allows us to obtain $m_{A_i \cap B_j}$ as the PRODUCT of $m_{A_i}$ and $m_{B_j}$. This has been discussed in section 3.5.3.

   (b) Detection-type fuzzy sets: $P(x, y|A_i \cap B_j) = P(x|A_i)P(y|B_j)$.

   Observations $x$ and $y$ are conditionally independent given the fuzzy sets that describe them. In this case we can obtain $m_{A_i \cap B_j}(x, y)$ from $P(x|A_i)$, $P(y|B_j)$ ("likelihood probabilities") and $P(A_i \cap B_j)$ (prior probability). This happens for detection-type fuzzy sets that act as underlying hypotheses, and has been discussed in section 3.6. In this case $P(A_i \cap B_j)$ carries
the information about the relation between $A_i$, $B_j$.

3.8 Probability and Fuzzy Sets: Zadeh’s View, Our View

In “Probabilistic Measures of Fuzzy Events” [Zad 68], Zadeh dealt with a probability measure defined on the observation space $U$ (or the “universe of discourse” fuzzy sets are defined on). Zadeh took $U$ as the sample space. He defined “fuzzy events”, a new type of events, as fuzzy sets in $U$. We look at some of his results in the following, and compare with our results.

1. Zadeh generalized the idea of an event (a crisp set of sample points) to that of a fuzzy event (a fuzzy set of sample points) as follows:

   **Zadeh’s Definition:** Let $(U, F, P)$ be a probability space in which $F$ is the $\sigma$-field of Borel sets in $U$ and $P$ is a probability measure over $U$. Then, a fuzzy event in $U$ is a fuzzy set $A$ in $U$ whose membership function, $m_A$ ($m_A: U \rightarrow [0, 1]$), is Bore measurable.

2. Our view: the probability space should be $(S, F', P')$, where $F'$ is the $\sigma$-field of Borel sets in the sample space $S$. $S$ is obtained by augmenting $U$, e.g. $S = U \times L$, where $U$ is the observation space, and $L$ is the description space as defined in section 3.3. $P'$ is a probability measure defined on $S$. A fuzzy set in $U$ is defined as an event (i.e. a crisp set) in $S$.

- Zadeh defined the probability of a fuzzy event $A$ by the following Lebesgue-Stieltjes integral:

$$ P(A) = \int_U m_A(x) \, dP = E(m_A). $$

---

*In Zadeh’s work, the observation space is taken as $R^n$. We replace it with $U$ to be consistent with our previous notation.*
• Our result agrees with Zadeh's. The probability of fuzzy set \( A \) (an event in \( S \)) is

\[
P(A) = \int_U P(A|x) dP = \int_U m_A(x) dP.
\]

• Zadeh also defined the **mean** and **variance** of a fuzzy event as follows.

1. The **mean** of a fuzzy event \( A \) relative to the probability measure \( P \) is:

\[
mm_P(A) = \frac{1}{P(A)} \int_U x m_A(x) dP.
\]

2. The **variance** of a fuzzy event \( A \) is:

\[
G_P^2(A) = \frac{1}{P(A)} \int_U (x - mm_P(A))^2 m_A(x) dP.
\]

• Our results agree with Zadeh's, but in our view,

1. what Zadeh calls the **mean** of fuzzy event \( A \), we would call the conditional mean of \( X \) given event (fuzzy set) \( A \), \( E(X|A) \):

\[
mm_P(A) = \int_U x \frac{m_A(x)}{P(A)} dP
\]

\[
= \int_U x \frac{P(A|x)}{P(A)} dP
\]

\[
= \int_U x dP(x|A)
\]

\[
= E(X|A).
\]

2. Similarly, what Zadeh calls the **variance** of a fuzzy event \( A \) we would call the conditional variance of \( X \) conditional on \( A \):

\[
G_P^2(A) = \int_U (x - mm_P(A))^2 \frac{m_A(x)}{P(A)} dP
\]

\[
= \int_U (x - E(X|A))^2 \frac{P(A|x)P(x)}{P(A)} dx
\]

\[
= \int_U (x - E(X|A))^2 dP(x|A)
\]

\[
= Var(X|A)
\]
• Zadeh defined the entropy of a fuzzy set $A$ as (assuming $x$ takes on finite number of values):

$$H^p(A) = -\sum_{i=1}^{n} m_A(x_i)P(x_i)\log P(x_i).$$

• Our result disagrees with Zadeh’s. $A$ is an event, but entropy is defined for a probabilistic distribution. We can define the entropy of $X$ conditional on $A$ based on the pseudomembership function ($pm_A(x_i) = P(x_i|A)$) as:

$$H^p(A) = -\sum_{i=1}^{n} P(x_i|A)\log P(x_i|A)$$

$$= -\sum_{i=1}^{n} \frac{m_A(x_i)P(x_i)}{P(A)} \log \frac{m_A(x_i)P(x_i)}{P(A)}$$

3.9 Conclusions

We conclude that the mathematical structure of probability can strengthen and enrich fuzzy set theory. Probability provides the necessary tools to take into account relations between fuzzy sets when combining multiple fuzzy sets. Many problems in conventional fuzzy set theory are naturally resolved. Conventional fuzzy set theory is not discarded, but put into the new perspective of an extended framework.

• Fuzzy sets can be viewed as events (crisp sets) in the enlarged sample space. The enlarged sample space can be viewed as the cartesian product of the observation space and the description space. New ways to visualize fuzzy sets are suggested.

• Fuzzy sets can be related in the same way as crisp sets are, with more varieties in componentwise relations. Typical relations encountered in practice are discussed.

• Union and intersection of fuzzy sets are taken based on the relations between fuzzy sets.
• All basic laws of logic can be satisfied by fuzzy sets, including the law of excluded middle and the law of contradiction, without sacrificing idempotency or distributivity.

• Probability suggests that there are two types of fuzzy sets. For quantization-type fuzzy sets, fuzzy sets act as descriptions, and membership functions are more natural. For detection-type fuzzy sets, fuzzy sets act as underlying hypotheses, and pseudomembership functions are more natural.

• Probability provides a unifying framework that allows us to interpret fuzzy sets probabilistically and to interpret random phenomena with fuzzy sets.

With these we move on to fuzzy control and fuzzy pattern recognition in the next two chapters, where we investigate the influence of the probabilistic interpretations in these applications.
Chapter 4

Fuzzy Control

4.1 Introduction

In this chapter we look at the impact of probabilistic interpretations on fuzzy control. For a more detailed introduction to fuzzy control and implementation issues, see [Lee 90][Wan 93][Jan 92][Jan 93].

The general structure of a control system is shown in Fig. 4.1.

![Diagram of a control system]

Figure 4-1: General structure of a control system.
The control system consists of the plant to be controlled and the controller that exerts an appropriate control action. The system dynamics in traditional control theory are modeled by a set of differential or difference equations, which tells how the system state responds to the control action applied by the actuator. A cost is defined as a function of the system states and the control actions. This cost is minimized by the optimal control function, which the controller attempts to implement. The traditional control system design starts with building a mathematical model that describes the system, defining the cost, and finding the optimal control function that minimizes this cost (and is stable). It can be difficult to find the differential/difference equations for the system model. Even if they are available, finding the optimal control function can still be difficult, especially when there is nonlinearity.

The design philosophy for fuzzy controllers is rather different. Such a design requires an expert who knows how to control the system. The way the expert controls the system can be viewed as an "ideal control function". The "ideal control function" might not be known as a numerical function, but information about it is available either in terms of linguistic descriptions (e.g. "if input is small then output is large"), or as numerical input/output training samples. The goal of the fuzzy controller is to implement this "ideal control function" through the use of fuzzy sets (to model descriptions such as "small", "large") and linguistic fuzzy if-then rules.

In a sense, fuzzy controllers do not solve the same problem as conventional optimal controllers. The more difficult task of finding "how to control" is done by the expert. The information about the "ideal control function" is always available in some form. As a result, most fuzzy controllers can be viewed as function approximators. The fuzzy controller implements an approximation of the "ideal control function" provided by the expert, rather than an optimal control function found from the mathematical model. The major effort in fuzzy controller design is to obtain a good approximation.

The distinctive feature of fuzzy controllers is that the functions they implement are described by local linguistic rules that are easy for human to generate, understand, and modify. Fuzzy sets model linguistic descriptions such as "small" and "large" better than crisp sets. The human-friendly local linguistic descriptions and the simple
structure are major advantages of fuzzy controllers, especially when the control function is simple enough to be thoroughly understood and linguistically described by an expert.

Chapter Outline

Section 4.2 is a summary of Tomizuka\cite{Tom 93}'s article that discusses issues of general interest about fuzzy control, e.g. the role of model, stability, the contribution of fuzzy control, and so on.

In section 4.3 we introduce the structure of a fuzzy controller. The basic operations of the fuzzy controller (fuzzification, inference, composition, and defuzzification) are discussed.

In section 4.4 we present the probabilistic interpretation of a fuzzy controller, and how the probabilistic view affects the design of fuzzy controllers. The probabilistic interpretation suggests using product fuzzification, product inference, and sum composition, and also suggest that input fuzzy sets should form a fuzzy partition, and output fuzzy membership functions should be replaced by pseudomembership functions.

In section 4.5 we study fuzzy controllers as function approximators. The trade off between linguistic meaning and numerical efficiency is discussed. The convergence rate of a fuzzy controller in terms of the number of fuzzy if-then rules and the order of input/output fuzzy membership functions is given.

4.2 Fuzzy control in Control Engineering

Tomizuka\cite{Tom 93}'s article provides much insight in various aspects of fuzzy controllers from the perspective of control engineering. His views are summarized as follows.
Role of models in the control system design.

The traditional approach to the design of control systems starts with mathematical models which describe the linear or nonlinear dynamics of the plant. The fuzzy design approach bypasses mathematical modeling and analysis, because the control algorithm is developed based on the experience and knowledge of experts such as plant operators, engineers, and plant designers.

The designers must have an implicit model, although it may not be in the differential equation form. Many successful fuzzy controllers have been designed without explicit models such as differential equations. This does not imply that the design is accomplished without a model, nor that it is a backward move to use explicit models in the design of fuzzy rule base controllers.

Stability.

The lack of stability assurance is one criticism against fuzzy controllers. There have been attempts to assure stability for various fuzzy controllers. Such stability arguments are made, naturally, for a limited class of systems; an explicit model is required, and techniques such as the Lyapunov stability and the small gain theorem are useful tools in those studies.

Note, however, that the lack of stability assurance is not unique in fuzzy control. Any control theory is based on a set of assumptions, and whether the assumptions are satisfied in application of the theory is another matter. Extensive testing is normal practice in control system design as the final check for stability and performance.

What fuzzy control brought to control engineering.

The fuzzy approach is based on engineering insights rather than mathematical modeling; it provides a friendly interface to humans. As long as the dynamics are simple, the human can control systems with complicated behaviors, incorporating many forms of feedback and feedforward information; wind surfing is a good example.

Engineers, without formal training in control theory, may design the controller
based on their physical understanding of the plant. The use of fuzzy logic is not limited to signal regulation level. Fuzzy logic can be used for higher level controls such as coordination of various local controllers and translation of symbols to variables in intelligent control systems.

**Blending conventional control and fuzzy control.**

If an explicit model is available, and a specific control theory is identified to be appropriate to solve a problem, the theory will provide an excellent initial design. If the design does not satisfy every requirement or causes something unexpected, a fuzzy rule based approach can be used for fine tuning.

When engineering insights make the fuzzy approach a more direct method, fuzzy control may be a primary tool. Even in such a situation, another control theory may be a viable tool for modifying or taming the plant to make it easier for the development of rules, especially when the original dynamics of the plant are very complex.

### 4.3 Structure of a Fuzzy Controller

#### 4.3.1 Pure Fuzzy Controllers

Consider a fuzzy controller with two input variables $x_1$ and $x_2$, one output variable $y$ ($x_1, x_2, y$ are real numbers), and fuzzy sets $A_j, B_j, C_j$ defined for $x_1, x_2,$ and $y$, respectively. There are $r$ linguistic fuzzy if-then rules, and each rule corresponds to a triplet of fuzzy sets $(A_j, B_j; C_j)$, $j = 1 \ldots r$.

Fuzzy controllers where both the input and output variables of the if-then rules are described by fuzzy sets are called "pure fuzzy controllers". They are the most common fuzzy controllers in the literature. An alternative is the Sugeno type fuzzy controller[Sug 85], which will be discussed in Section 4.2.2.

The rules for the pure fuzzy controller are of the form:

---

1. Fuzzy controllers with multiple output variables can always be thought of as several fuzzy controllers each with a single output.
If \( x_1 \) is \( A_1 \) and \( x_2 \) is \( B_1 \) then \( y \) is \( C_1 \)

If \( x_1 \) is \( A_2 \) and \( x_2 \) is \( B_2 \) then \( y \) is \( C_2 \)

....

We describes what these rules actually mean in what follows.

For example, a rule can be:

If \( x_1 \) is Large and \( x_2 \) is Small then \( y \) is Medium.

For any input \((x_1, x_2)\), each rule is activated to some degree, depending on how well \( x_1 \) and \( x_2 \) agree with the descriptions in the IF part of the rule.

Let \( x \) be the vector of input variables; here \( x = (x_1, x_2) \). When the antecedent part of each rule contains conditions on several variables connected with "AND", an input fuzzy set \( I_j \), can be defined as the intersection of fuzzy sets describing each variable. In our example the \( j \)th input fuzzy set is

\[
I_j = A_j \cap B_j.
\]

A word of caution about the index: For convenience fuzzy sets \( A_j, B_j \) are indexed by the rule they belong to, so it is possible to have \( A_i = A_j \) but \( i \neq j \). See the following example.

**Example:** Suppose there are two fuzzy sets "small", "large" defined over each input variable \( x_1, x_2 \). \( S^1, L^1 \) are defined over \( x_1 \), and \( S^2, L^2 \) are defined over \( x_2 \). There will be four input fuzzy sets, \( I_1, \ldots I_4 \).

\[
I_1 = A_1 \cap B_1 = S^1 \cap S^2 \\
I_2 = A_2 \cap B_2 = L^1 \cap S^2 \\
I_3 = A_3 \cap B_3 = S^1 \cap L^2 \\
I_4 = A_4 \cap B_4 = L^1 \cap L^2
\]

Note that \( B_1 = B_2 = S^2, A_1 = A_3 = S^1 \).

In what follows we will introduce how the fuzzy controller works, that is, how \( y \)
is obtained from $x_1, x_2$ through the fuzzy sets and the fuzzy if-then rules.

The main operations of a pure fuzzy controller are: Fuzzification, inference, composition, and defuzzification. (see Fig. 4.2).

![Diagram of fuzzy controller]

**Figure 4-2: General structure of a fuzzy controller.**

1. **FUZZIFICATION**: For given inputs $x_1, x_2$, find the membership values $m_{A_j}(x_1)$ and $m_{B_j}(x_2)$ for each rule, $j = 1 \ldots r$. Combine these values to obtain the **firing strength** for each rule $j$ (i.e. the degree to which rule $j$ is activated).

The firing strength for rule $j$ is the membership function value of the input fuzzy set $I_j$ at $x = (x_1, x_2)$. That is,

$$firing\ strength = m_{I_j}(x) = m_{A_j \cap B_j}(x_1, x_2).$$

In the literature, firing strength is obtained by applying fuzzy AND operations such as MIN, PRODUCT, or other operations.

$$m_{I_j}(x) = \min[m_{A_j}(x_1), m_{B_j}(x_2)]$$

or

$$m_{I_j}(x) = m_{A_j}(x_1)m_{B_j}(x_2)$$

2. **INFEERENCE**: For each rule $j$, the firing strength $m_{I_j}(x)$ is applied to the output fuzzy set $C_j$ to produce an “activated output fuzzy set” $O_j$. The common inference rules are MIN and PRODUCT.
In **MIN inferencing**, the output membership function is “clipped off” at a height that is equal to the “firing strength”.

\[ m_{O_j}(y) = \min[m_{I_j}(x), m_{C_j}(y)] \]

Note that the fuzzy set \( O_j \) is defined on the observation space of \( y \), but depends on the particular input \( x \).

In **PRODUCT inferencing**, the output membership function is scaled (multiplied) by the firing strength.

\[ m_{O_j}(y) = m_{I_j}(x)m_{C_j}(y) \]

3. **COMPOSITION**: The “activated output fuzzy sets” \( O_j \) obtained for all rules are combined into one “activated overall output fuzzy set”, \( O \), which can be viewed as the output. The common composition rules are **MAX** and **SUM**. That is, \( m_O(y) \) is the pointwise MAX or SUM of \( m_{O_j}(y) \), \( j = 1, \ldots r \).

\[ m_O(y) = \max_j m_{O_j}(y) \]

or

\[ m_O(y) = \sum_{j=1}^{r} m_{O_j}(y) \]

4. **DEFUZZIFICATION**: Convert the “activated overall output fuzzy set” \( O \) into a crisp number \( y_{out} \), the defuzzified output. There are many ways to defuzzify. The most common ones are **CENTROID** and **MAX**.

In **CENTROID defuzzification**, the defuzzified output value is the “center of gravity” of the overall output fuzzy set. \(^2\)

\(^2\)\( y \) is assumed to be discrete here. When \( y \) is continuous, replace the summation by integration over \( y \).
\[ y_{out} = \frac{\sum y m_o(y) y}{\sum y m_o(y)} \]

In **MAX defuzzification**, one of the values where the overall output fuzzy membership function achieves its maximum is chosen as the output value.

\[ y_{out} = \arg \max_y m_o(y) \]

When product fuzzification, product inference, sum composition and centroid defuzzification are used, we will refer to the fuzzy controller as a **PPSC fuzzy controller**. In this case the output \( y_{out} \) is the convex combination of the centroids of \( C_j, j = 1, \ldots, r \).

\[
y_{out} = \frac{\sum_{j=1}^{r} \sum y m_{I_k}(x) m_{C_j}(y) y}{\sum_{k=1}^{r} \sum y m_{I_k}(x) m_{C_k}(y)} \]

\[ = \frac{\sum_{j=1}^{r} m_{I_k}(x) \sum y m_{C_j}(y) y}{\sum_{k=1}^{r} m_{I_k}(x) \sum y m_{C_k}(y)} \]

\[ = \sum_{j=1}^{r} S_j(x) \mu_{C_j} \]

Where

\[ \mu_{C_j} = \frac{\sum y m_{C_j}(y)}{\sum y m_{C_j}(y)} \]

and

\[ S_j(x) = \frac{m_{I_k}(x) \sum y m_{C_j}(y)}{\sum_k m_{I_k}(x) \sum y m_{C_k}(y)} \]

Fuzzy if-then rules for PPSC fuzzy controllers can be written as

If \( x \) is \( S_j \) then \( y \) is \( \mu_{C_j} \)

instead of “If \( x \) is \( I_j \) then \( y \) is \( C_j \)”. The antecedent part of rule \( j \) corresponds to the normalized fuzzy set \( S_j \), and the consequent part corresponds to the centroid of fuzzy

---

\(^3\)A convex combination of \( \{\mu_{C_1}, \ldots, \mu_{C_r}\} \) is a weighted sum \( \sum_{j=1}^{r} p_j \mu_{C_j} \), where \( \sum_{j=1}^{r} p_j = 1 \) and each \( p_j \geq 0 \).
set $C_j$, $\mu_{C_j}$. Whatever the output membership function $m_{C_j}(y)$ may be (Gaussian, triangular, trapezoidal...), as long as $\sum_y m_{C_j}(y) = 1$ and the centroid $\mu_{C_j}$ remain the same, the fuzzy controller remains the same. In particular $C_j$ can be a singleton fuzzy set with $m_{C_j}(y) = 1$ when $y = \mu_{C_j}$ and $m_{C_j}(y) = 0$, otherwise.

### 4.3.2 Sugeno Fuzzy Controllers

For the Sugeno fuzzy controller [Sug 85], the consequent part of each fuzzy if-then rule corresponds not to a fuzzy set, but to a crisp number defined as the function of the input.

If $x_1$ is $A_1$ and $x_2$ is $B_1$ then $y = g_1(x_1, x_2)$

If $x_1$ is $A_2$ and $x_2$ is $B_2$ then $y = g_2(x_1, x_2)$

.....

The rules can also be written as:

If $x$ is $I_j$ then $y = g_j(x), j = 1 \ldots r$,

where $I_j = A_j \cap B_j$, and $x = (x_1, x_2)$.

The operations are a bit different from those of the pure fuzzy controllers. Under INERENCE, rule $j$ produces an output value (as opposed to an output fuzzy set) by taking the product of the firing strength with the function at the consequent part, $m_{I_j}(x)g_j(x)$, $j = 1, \ldots r$. Under COMPOSITION the final output value is obtained as a sum of the output value of each rule, normalized by the sum of the firing strengths.

$$y_{out} = \frac{\sum_{j=1}^r m_{I_j}(x)g_j(x)}{\sum_{j=1}^r m_{I_j}(x)}.$$

The pure PPSC fuzzy controller (product fuzzification, product inference, sum composition, centroid defuzzification) corresponds to a special case of Sugeno fuzzy controllers. Output of the pure PPSC fuzzy controller is:

$$y_{out} = \frac{\sum_{j=1}^r m_{I_j}(x)\sum_y m_{C_j}(y)\mu_{C_j}}{\sum_{j=1}^r m_{I_j}(x)\sum_y m_{C_j}(y)}.$$
This is the same as the Sugeno fuzzy controller where \( g_j(x) = \mu_{C_j} \) are constants that do not depend on \( x \), and \( \sum \mu_{C_j}(y) = 1 \). In this case the fuzzy if-then rules reduce to:

- If \( x_1 \) is \( A_1 \) and \( x_2 \) is \( B_1 \) then \( y \) is \( \mu_{C_1} \)
- If \( x_1 \) is \( A_2 \) and \( x_2 \) is \( B_2 \) then \( y \) is \( \mu_{C_2} \)

.....

As will be discussed in Section 4.5, the roles of the input fuzzy membership functions and the output functions are very similar. Note that although pure fuzzy controllers appear to be a special case of the Sugeno type fuzzy controllers, no logical generality is introduced by the Sugeno type fuzzy controllers. Any input/output mapping can be achieved by either a Sugeno fuzzy controller (by selecting particular output functions) or a pure fuzzy controller (by selecting particular input fuzzy membership functions); see section 4.5.

### 4.4 Probabilistic Interpretations of a Fuzzy Controller

#### 4.4.1 Pure Fuzzy Controllers

In the probabilistic view, the input vector \( x \) can be viewed as a vector random variable \( X \), and the intermediate fuzzy sets \( A_j, B_j, C_j \) can all be modeled as being random. Given \( X = (x_1, x_2) \), the conceptual output can also be viewed as a random variable \( Y \), defined in terms of \( x \) by \( P_{Y|X}(y|x) \). The defuzzified output \( y_{out} \) is a "representative value", i.e. a statistic determined from \( P_{Y|X}(y|x) \) by defuzzification.

The operations of the fuzzy controller can be modeled as events "occurring" in sequence. There is a Markov property between them.

Equivalently, with \( X = (X_1, X_2) \), \( I_j = A_j \cap B_j \), we have

\[
X \rightarrow I_j \rightarrow C_j \rightarrow Y.
\]
The Markov property means that:

\[ P(y|C_j, I_j, x) = P(y|C_j) \]

\[ P(C_j|I_j, x) = P(C_j|I_j). \]

We make the following assumptions:

- The fuzzy sets \( I_j = A_j \cap B_j \), \( j = 1 \ldots r \) partition the enlarged input sample space. As a result, \( \sum_{j=1}^{r} m_{I_j}(x) = 1 \).

- Given \( I_j \), the fuzzy set \( C_j \) occurs with probability one. Therefore, \( C_j, j = 1, \ldots r \) partition the enlarged output sample space. Fuzzy set \( C_j \) is viewed as a detection-type fuzzy set with pseudomembership function \( \sum_{y} pm_{C_j}(y) = 1 \).

Existing fuzzy controllers turn out to be equivalent to fuzzy controllers that satisfy these assumptions. This will be explained later in this section.

Events "occurring" in the fuzzy controller can be visualized through a probability tree. See Fig. 4.3.

In the fuzzy model, given \( x \) all \( I_j \)'s "occur" simultaneously, each to some degree. In the probabilistic model, only one randomly chosen \( I_j \) occurs, even though the final output of the system is a statistic of the random model, and is perfectly deterministic. Using probability as a formal model proves to be rather useful, however. We gain the ability to deal with relations between fuzzy sets, and there are standard ways to take union/intersections according to the relations. Seemingly arbitrary choice in membership functions, operations for fuzzification, inference and composition now have a natural interpretation.
Figure 4-3: Pure Fuzzy Controller: a Probability Tree Representation.

We will explain in the following how the sequence of events "occur", and what is the corresponding probabilistic interpretation of fuzzification, inference, composition and defuzzification. See Fig. 4.4.

1. FUZZIFICATION.

- Fuzzy View:
  Given the input $x$, find the firing strengths $m_{I_j}(x) = m_{A_j \cap B_j}(x_1, x_2)$, $j = 1, \ldots r$.

- Probability View:
  Given $x$, it is in fuzzy set $I_j$ to some degree. This is modeled as the event (fuzzy set) $I_j$ "occurs" with probability $P(I_j|x) = P(A_j \cap B_j|x_1, x_2)$. $A_j$ only describes $x_1$ and $B_j$ only describes $x_2$, so they can be modeled as being conditionally independent (see section 3.5.2). The membership function of $A_j \cap B_j$ is thus the product of the individual membership functions.
Figure 4-4: Probabilistic View of a Fuzzy Controller.

\[
m_{I_j}(x) = P(I_j | x) \\
= P(A_j \cap B_j | x_1, x_2) \\
= P(A_j | x_1) P(B_j | x_2) \\
= m_{A_j}(x_1) m_{B_j}(x_2)
\]

In other words, probability suggests using **PRODUCT** in fuzzification to obtain the firing strength \( m_{I_j}(x) = m_{A_j \cap B_j}(x_1, x_2) \), due to the conditional independence of fuzzy sets \( A_j \) and \( B_j \).

2. INFERENCE.

- **Fuzzy View:**
  Obtain an output fuzzy membership function \( m_{O_j}(y) \) for rule \( j \) by combining the firing strength \( m_{I_j}(x) \) and the consequent fuzzy membership function \( m_{C_j}(x_1, x_2) \).

- **Probability View:**
  Given \( A_j, B_j \), the event \( C_j \) occurs with probability one. Given \( C_j \), the event \( Y = y \) occurs with probability \( P(y | C_j) \). This conditional probability
corresponds to the pseudomembership function of fuzzy set \( C_j \), \( pm_{C_j}(y) = P(y|C_j) \) (see section 3.6). In the probability view, the fuzzy sets \( C_j \), \( j = 1, \ldots, r \) are given as underlying hypotheses to generate the output value \( y \). It is thus the pseudomembership function that we use to generate \( y \).

The “output fuzzy membership function” \( m_{O_j}(y) \) corresponds to the probability \( P(y, C_j, A_j, B_j|x_1, x_2) \). In the probabilistic view it relates more to a pseudomembership function over \( y \) than a membership function. Due to the Markov property, it can be factored as follows.

\[
P(y, C_j, A_j, B_j|x_1, x_2) = P(y|C_j, A_j, B_j, x_1, x_2)P(C_j, A_j, B_j|x_1, x_2) \\
= P(y|C_j)P(C_j|A_j, B_j)P(A_j, B_j|x_1, x_2) \\
= P(y|C_j)P(A_j|x_2)P(B_j|x_2) \\
= pm_{C_j}(y)[m_{A_j}(x_1)m_{B_j}(x_2)] \\
= pm_{C_j}(y)m_{I_j}(x)
\]

Probability thus suggests using **PRODUCT INFERENCE**. It also suggests that \( m_{C_j}(y) \) should be replaced by the pseudomembership functions \( pm_{C_j}(y) = P(y|C_j) \). That is, for fuzzy sets \( C_j \) we should define pseudomembership functions that add up to one over \( y \) as opposed to membership functions. This is of course satisfied if the output fuzzy sets are singleton fuzzy sets (\( m_{C_j}(y) = 1 \) for \( y = \mu_{C_j} \), \( m_{C_j}(y) = 0 \) otherwise).

Note that with the probability interpretation, from Baye’s rule it is required that:

\[
pm_{C_j}(y)P(C_j) = m_{C_j}(y)P(y).
\]

From the probabilistic view one cannot obtain pseudomembership func-
tions from membership functions simply by normalization:

\[ \text{pm}_{c_j}(y) \neq \frac{m_{c_j}(y)}{\sum_y m_{c_j}(y)}, \]

unless \( P(y) \) is uniform over \( y \). Defining the membership functions is different from defining the pseudomembership functions. Whether membership functions or pseudomembership functions are more appropriate depends on the nature of the process (see Section 3.6).

3. COMPOSITION.

- **Fuzzy View:**
  Combining \( m_{o_j}(y), j = 1, \ldots r \) to obtain the overall output fuzzy membership function \( m_{o}(y) \).

- **Probability View:**
  Combining \( P(y, c_j, a_j, b_j|x_1, x_2), j = 1, \ldots r \) to obtain the overall output pseudomembership function \( \text{pm}_{o}(y) = P(y|x_1, x_2) \).

\[
\text{pm}_{o}(y) = P(y|x_1, x_2) \\
= \sum_j P(y, c_j, a_j, b_j|x_1, x_2) \\
= \sum_j \text{pm}_{c_j}(y)m_{i_j}(x)
\]

Probability thus suggests **SUM COMPOSITION**. The overall output pseudomembership function is the convex combination (interpolation) of the output pseudomembership functions \( \{\text{pm}_{c_j}(y)\} \).

4. DEFUZZIFICATION.

- **Fuzzy View:** Obtain a crisp number \( y_{out} \) from the overall output fuzzy set with membership function \( m_{o}(y) \).

- **Probability View:** Obtain a representative number \( y_{out} \) from \( \text{pm}_{o}(y) = P(y|x_1, x_2) \), the pseudomembership function of the overall output fuzzy
set. Probability does not suggest any particular defuzzification method. The "representative value" of a random variable \((y \text{ given } x_1, x_2)\) can be its mean (centroid defuzzification), its mode (maximum defuzzification) or other choices.

**CENTROID DEFUZZIFICATION**

\[
y_{out} = \sum_y y p_m(y)
\]

\[
= \sum_y y P(y|x_1, x_2)
\]

\[
= E(Y|x_1, x_2)
\]

**MAXIMUM DEFUZZIFICATION**

\[
y_{out} = \arg\max_y \, p_m(y)
\]

\[
= \arg\max_y \, P(y|x_1, x_2)
\]

It is interesting to note that, when we view finding \(y_{out}\) as estimating \(Y\) given observation \(x_1, x_2\) with a fixed \(P(y|x_1, x_2)\), centroid defuzzification gives the conditional mean \(y_{out} = E(Y|x_1, x_2)\), which is the Bayesian least square estimate in the sense that it minimizes the "mean squared error" \(E[(Y - y_{out})^2]\). Maximum defuzzification gives the aposteriori mode \(y_{out} = \arg\max_y \, P(y|x_1, x_2)\), which is the MAP estimate (maximum aposteriori probability estimate).

In the design of a fuzzy controller, however, the transition probability \(P_{Y|X}(y|x)\) is not fixed initially. A major design effort is in finding the best transition probability \(P_{Y|X}(y|x)\), which corresponds to finding the fuzzy sets and the fuzzy if-then rules. The fuzzy sets and the rules are more important than the defuzzification method in determining \(y_{out}\) as a function of \(x\). The "optimality" of centroid defuzzification can be meaningless if \(P_{Y|X}(y|x)\) is poorly chosen. On the other hand, whatever defuzzification method is used, \(y_{out}\) as a function of \(x\) can be easily changed by tuning the fuzzy sets and the fuzzy if-then rules, and thus \(P_{Y|X}(y|x)\). The same argument
applies to fuzzification, inference and composition. Their roles are not as significant as those of the fuzzy sets and the rules in terms of affecting $y_{out}$ as a function of the input.

One might notice that fuzzy controllers in practice do not seem to satisfy all the requirements for probabilistic interpretations.

- The input fuzzy membership functions do not always add up to one, and thus do not form a fuzzy partition:

$$\sum_j m_{I_j}(x) = \sum_j m_{A_j \cap B_j}(x_1, x_2) = \sum_j m_{A_j}(x_1)m_{B_j}(x_2) \neq 1.$$ 

- Probability suggests using the pseudomembership functions $pm_{C_j}(y) = P(y|C_j)$ for the output fuzzy sets. The membership functions $m_{C_j}(y), j = 1, \ldots, r,$ used in existing fuzzy controllers do not add up (or integrate) to one over $y$ in general.

$$\sum_y m_{C_j}(y) \neq 1.$$ 

However, we now show that due to the normalization in centroid defuzzification, these requirements are satisfied indirectly in a conventional PPSC fuzzy controller with product fuzzification, product inference, sum composition, and centroid defuzzification. The PPSC fuzzy controllers can always be viewed probabilistically as producing a conditional mean. For a PPSC fuzzy controller, the output is

$$y_{out} = \frac{\sum_j \sum_y m_{A_j}(x_1)m_{B_j}(x_2)m_{C_j}(y)y}{\sum_j \sum_y m_{A_j}(x_1)m_{B_j}(x_2)m_{C_j}(y)}.$$ 

Let $\alpha_j = [\sum_y m_{C_j}(y)]^{-1}$. We define fuzzy set $C'_j$ to have pseudomembership function

$$pm_{C'_j}(y) = P(y|C'_j) = \alpha_j m_{C_j}(y).$$

Note that

$$\sum_y pm_{C'_j}(y) = 1.$$
Let $m_{A_i'}(x_1)m_{B_i'}(x_2) = (\alpha_j)^{-1}m_{A_i}(x_1)m_{B_i}(x_2)$. Replace $m_{A_i}(x_1)m_{B_i}(x_2)m_{C_j}(y)$ by $m_{A_i'}(x_1)m_{B_i'}(x_2)m_{C_j}(y)$ in both the numerator and denominator of $y_{out}$. Then we have

$$
y_{out} = \frac{\sum_j \sum_y m_{A_i'}(x_1)m_{B_i'}(x_2)p_{m_{C_j}}(y)y}{\sum_j \sum_y m_{A_i}(x_1)m_{B_i}(x_2)p_{m_{C_j}}(y)}.$$

Next choose $\beta(x_1, x_2) = [\sum_j m_{A_i'}(x_1)m_{B_i'}(x_2)]^{-1}$.

Let $m_{A_i'}(x_1)m_{B_i'}(x_2) = \beta(x_1, x_2)[m_{A_i'}(x_1)m_{B_i'}(x_2)]$. Now

$$\sum_j m_{A_i''}(x_1)m_{B_i''}(x_2) = 1.$$

Multiplying both the numerator and the denominator of $y_{out}$ by $\beta(x_1, x_2)$, the denominator is reduced to one, and we have

$$y_{out} = \sum_j m_{A_i''}(x_1)m_{B_i''}(x_2)\sum_y p_{m_{C_j}}(y)y = \sum_j P(A_{i''}, B_{i''}|x_1, x_2)\sum_y P(y|C_j)\gamma = \sum_j P(C_{j'}|x_1, x_2)E(y|C_{j'}) = E(Y|x_1, x_2)$$

The Markov property here is:

$$X_1 \rightarrow A_j \rightarrow C_j \rightarrow Y$$

$$X_2 \rightarrow B_j \rightarrow Y$$

90
Viewing \( y_{out} = \sum_{j} P(C'_j|x_1, x_2)E(Y|C'_j) \) shows that it is only the centroids \( \mu_{C'_j} = E(Y|C'_j) \) that affect the fuzzy controller. Since \( y_{out} \) is a convex combination of the centroids of \( C'_j \), the maximum/minimum values \( y_{out} \) can take is equal to the maximum/minimum values of the centroids.

Alternatively, note that \( P(C'_j|A''_j, B''_j) = 1 \), and \( P(y|x_1, x_2) = \sum_j P(y, C'_j|x_1, x_2) \).

We can thus view \( y_{out} \) in another way:

\[
y_{out} = \sum_j m_{A''_j}(x_1)m_{B''_j}(x_2) \sum_y p m_{C'_j}(y)y \\
= \sum_j \sum_y P(A''_j, B''_j|x_1, x_2)P(y|C'_j)y \\
= \sum_y \sum_j P(y|C'_j)P(C'_j|x_1, x_2)y \\
= \sum_y P(y|x_1, x_2)y \\
= E(Y|x_1, x_2).
\]

So, for any fuzzy controller with product firing strength, product inference, sum composition, and centroid defuzzification:

- The fuzzy sets \( A_j, B_j, C_j \) used in the fuzzy controller can be replaced by normalized fuzzy sets \( A''_j, B''_j, C'_j \) without changing the fuzzy controller. \( A''_j, B''_j, C'_j \) conform to the probabilistic interpretations.

- The output \( y_{out} \) can always be viewed as a conditional mean \( E(Y|x_1, x_2) \), whether or not fuzzy sets specified by the designer satisfy the probabilistic requirements.

The above seem to suggest that one need not worry about whether the fuzzy sets are normalized, as the normalization will be taken care of in defuzzification. However, the normalized fuzzy sets can look very different from the fuzzy sets one originally specified. One should recognize that it is the normalized fuzzy sets that really matter. Forgetting this can cause misunderstanding, as shown in the following.

For example, the concern of "enough coverage" for input fuzzy sets is unnecessary,
as long as there is any coverage at all. The “coverage” is defined as the sum or maximum of all the input membership function values at each point. It is thought that the coverage should be greater than some positive constant over the whole range.

To see that this is unnecessary, consider a PPSC fuzzy controller with one input $x$ and input fuzzy sets $A_j$, $j = 1, \ldots, r$. See Fig. 4.5. This is a special case of Sugeno controller where the rules are:

If $x$ is $A_j$ then $y$ is $\mu_j$, $j = 1 \ldots r$.

Here

$$y_{out} = \frac{\sum_j m_{A_j}(x)\mu_j}{\sum_j m_{A_j}(x)}.$$ 

Using input fuzzy sets $A_j$ on the left of Fig. 4.5 is exactly the same as using the normalized fuzzy sets $A'_j$ on the right, where $m_{A'_j}(x) = m_A(x)/\sum_j m_{A_j}(x)$. This can be seen from the above expression for $y_{out}$. There will always be enough coverage, as long as there is any coverage at all, due to the normalization. If there is no coverage in some regions of $x$ ($\sum_j m_{A_j}(x) = 0$), the output $y_{out}$ is undefined when the input $x$ falls into those regions.

A popular choice for fuzzy membership functions is the Gaussian function.

$$m_{A_j}(x) = \exp\left(-\frac{(x - \mu_{A_j})^2}{2\sigma_{A_j}^2}\right)$$

In this case it is easier to specify the mean and variance of each Gaussian function.
first and normalize them later.

\[ m_{A_i}(x) = \frac{m_{A_j}(x)}{\sum_j m_{A_j}(x)} \]

One should realize that the normalized membership functions \( m_{A_i}(x) \), the ones that matter, are no longer Gaussian. They might look rather different from the initially specified Gaussian functions (see Fig. 4.6).

![Figure 4-6: Gaussian-type fuzzy membership functions.](image)

In summary, probability suggests always using normalized fuzzy sets. This requires:

- Defining pseudomembership functions, as opposed to membership functions, for output fuzzy sets. These sum (integrate) to one over possible output values.

- Defining input fuzzy sets for which the membership functions add up to one over all fuzzy sets (form a fuzzy partition). When it is more convenient to define initial input fuzzy sets and normalize them later, remember it is the normalized input fuzzy sets that are eventually used in the fuzzy controller.
Different collections of input fuzzy sets that are the same after normalization will produce the same fuzzy controller.

Fuzzy Partition

If the fuzzy sets for each input variable form a partition, the combined input fuzzy sets using all possible combinations also form a partition. Consider the previous example where fuzzy sets $S^1$, $L^1$ are defined for $x_1$, fuzzy sets $S^2$, $L^2$ are defined for $x_2$. There are 4 totally input fuzzy sets.

\[
\begin{align*}
I_1 &= A_1 \cap B_1 = S^1 \cap S^2 \\
I_2 &= A_2 \cap B_2 = L^1 \cap S^2 \\
I_3 &= A_3 \cap B_3 = S^1 \cap L^2 \\
I_9 &= A_4 \cap B_4 = L^1 \cap L^2
\end{align*}
\]

When all 4 input fuzzy sets are used (4 fuzzy if-then rules), the requirement that they form a fuzzy partition:

\[
\sum_{j=1}^{4} m_{I_j}(x) = 1
\]

is satisfied if fuzzy sets for each input variable form a fuzzy partition,

\[
m_{S^1}(x) + m_{L^1}(x) = 1
\]

\[
m_{S^1}(x) + m_{L^1}(x) = 1.
\]

"Rule Deletion" and "Zero Output for Some Rules"

When there are more input variables and many fuzzy sets per variable, the number of rules can become very large. It is common in existing applications to reduce the number of rules. One way is to simply delete some of the rules [Yam 93]. Rule deletion (i.e. not using all possible input fuzzy sets) adds dependencies between input
fuzzy sets for different variables. For example, not using \( I_2 = L^1 \cap S^2 \) corresponds to having \( P(L^1 \cap S^2|X_1, X_2) = 0 \neq P(L^1|X_1)P(S^2|X_2) \) for all \( x_1, x_2 \). This leads to \( P(L^1 \cap S^2) = 0 \), which means that the input condition "\( x_1 \) is Large and \( x_2 \) is Small" never occurs (to any degree) for any \( x_1, x_2 \). This is contrary to our intuition and the probabilistic interpretation.

The argument for rule deletion given in [Yam 93] is that the controller does not need to take action for every possible input condition. As opposed to declaring that certain input conditions are not possible, this should be carried out by allowing all input conditions, but specifying the output as "no action" for certain fuzzy if-then rules (zero output).

Consider a PPSC fuzzy controller where the input fuzzy sets \( I_j \) form a partition, and the fuzzy if-then rules are:

\[
\text{If } x \text{ is } I_j \text{ then } y \text{ is } C_j, \ j = 1, 2, 3, 4.
\]

Not taking control action for input condition \( I_2 \) corresponds to \( \mu_{C_2} = 0 \). This can be achieved by having a singleton pseudomembership function \( pm_{C_2}(y) = 1 \) when \( y = 0 \) and \( pm_{C_2}(y) = 0 \) otherwise. The defuzzified output \( y_{def} \) given input \( x \) will be:

\[
y_{def} = \frac{\sum_{j=1}^{4} m_{I_j}(x)\mu_{C_j}}{\sum_{j=1}^{4} m_{I_j}(x)} = \sum_{j=1}^{4} m_{I_j}(x)\mu_{C_j} = m_{I_1}(x)\mu_{C_1} + m_{I_2}(x)\mu_{C_2} + m_{I_3}(x)\mu_{C_3} + m_{I_4}(x)\mu_{C_4} = m_{I_1}(x)\mu_{C_1} + m_{I_3}(x)\mu_{C_3} + m_{I_4}(x)\mu_{C_4}.
\]

Eliminating the rule associated with \( I_2 \) produces something different:

\[
y_{def} = \frac{m_{I_1}(x)\mu_{C_1} + m_{I_3}(x)\mu_{C_3} + m_{I_4}(x)\mu_{C_4}}{m_{I_1}(x) + m_{I_3}(x) + m_{I_4}(x)}.
\]

Note that \( y_{def} \) is scaled up, which presumably is not the intended effect. Operationally both equations omit \( m_{I_2}(x) \), but the first uses a normalization that corresponds to having no output for \( I_2 \) rather than assuming \( m_{I_2}(x) = 0 \).
4.4.2 Sugeno Fuzzy Controllers

The probabilistic interpretations for Sugeno fuzzy controllers are similar to those for pure fuzzy controllers.

In this case, the conceptual output can be viewed as a random variable $Y$ that takes on values $y_j = g_j(x_1, x_2)$ with probability $P(Y = y_j|x_1, x_2) = P(A_j \cap B_j|x_1, x_2)$, $j = 1 \ldots r$. The output $y_{out}$ corresponds to the conditional mean $E(Y|x_1, x_2)$.

\[
y_{out} = \sum_j m_{I_j}(x)g_j(x_1, x_2)
= \sum_j P(A_j \cap B_j|x_1, x_2)g_j(x_1, x_2)
= \sum_j m_{A_j}(x_1)m_{B_j}(x_2)g_j(x_1, x_2)
= \sum_j P(Y = y_j|x_1, x_2)y_j
= E(Y|x_1, x_2).
\]

The probabilistic interpretation suggests using product fuzzification to obtain the firing strength.

Just as for pure fuzzy controllers, in existing Sugeno fuzzy controllers the input fuzzy membership functions might not add up to one, and there is a normalization in the defuzzification:

\[
y_{out} = \frac{\sum_j m_{I_j}(x)g_j(x_1, x_2)}{\sum_j m_{I_j}(x)}
\]

It is just as important here to realize that it is the normalized input membership functions that matters, and it is better to think of the normalized functions as the input membership functions.
4.4.3 Cascading Two Fuzzy Controllers

The probabilistic interpretation provides insights into cascading fuzzy controllers. See Fig. 4.7.

The two fuzzy controllers are characterized by $P_{Y|X}$, $P_{Z|Y}$, respectively. There are two ways to cascade the fuzzy controllers.

- **Output of fuzzy controller 1 needs to be a crisp number.**
  Let the crisp output of fuzzy controller 1 (also the input to fuzzy controller 2) be $Y_o$. Given $x$, $Y_o$ will take on value $y_{def}$, which depends on $x$ and the defuzzification scheme. In this case the first fuzzy controller can be viewed as being characterized by a singleton probability $P_{Y_o|X}(y_o|x) = 1$ when $y_o = y_{def}$, and $P_{Y_o|X}(y_o|x) = 0$ otherwise. The combined fuzzy controller is characterized by $P_{Z|X}(z|x) = \int_{y_o} P_{Z|Y_o}(z|y_o)P_{Y_o|X}(y_o|x) = P_{Z|Y_o}(z|y_{def})$.

- **Output of fuzzy controller 1 needs not be a crisp number.**
  Note that the fuzzy if-then rules for the combined fuzzy controller can be obtained in a straightforward way from the if-then rules for the two component controllers.

The requirement is that every fuzzy set $B_j$ in the consequent part of rules in fuzzy controller 1 must be in the antecedent part of some rule in fuzzy controller 2. Suppose the rules for fuzzy controller 1 are:
If \( x \) is \( A_j \) then \( y \) is \( B_j \)

The rules for fuzzy controller 2 are:

If \( y \) is \( B_k \) then \( z \) is \( C_k \).

Then it is easy to match up the rules with the same fuzzy set over \( y \) and combine them into one single rule:

If \( x \) is \( A_j \) then \(( y \) is \( B_j \), if \( y \) is \( B_j \) then \) \( z \) is \( C_j \).

This can be combined into

If \( x \) is \( A_j \) then \( z \) is \( C_j \).

One can use the new set of rules directly to implement the combined controller as a single fuzzy controller with input \( X \) and output \( Z \). The number of rules will be the same as fuzzy controller 1.

Another possibility is, given \( x \), to send the vector of firing strengths \( \{ m_{A_j}(x) \} \) and use them to obtain the firing strengths for the corresponding rules in fuzzy controller 2. This will eliminate the inference, composition and defuzzification operations in fuzzy controller 1, and the fuzzification in fuzzy controller 2.

### 4.5 Fuzzy Controllers as Function Approximators

Fuzzy controllers can be viewed as function approximators. Information about the function to be approximated, \( f(x) \), might be in linguistic form (fuzzy if-then rules given by experts), or in numerical form (\( f(x) \) is a known numerical function or there are numerical input/output training samples). The fact that fuzzy controllers can approximate any continuous functions over a compact interval to arbitrary accuracy has been shown in [WaM 92] [Jan 93] [Bry 93].

In this section we study the performance of fuzzy controllers as function approximators. The “performance” is measured by the discrepancy (maximum error) between the function implemented by the fuzzy controller (call it \( g(x) \)) and the function it attempts to approximate (\( f(x) \)).
Define the maximum error between \( g(x) \) and \( f(x) \) as follows.

- **Maximum Error:**
  \[
  \max_{x \in U} |f(x) - g(x)|,
  \]

  where \( U = [0, 1] \) is the interval where the function approximation takes place.

  We will study Sugeno fuzzy controllers with rules:

  If \( x \) is \( A_j \) then \( y \) is \( g_j(x) \), \( j = 1 \ldots r \).

  Pure fuzzy controllers are a special case with \( g_j(x) = \mu_j \). The input membership functions are assumed to form a fuzzy partition, i.e. \( \sum_{j=1}^{r} m_{A_j}(x) = 1 \). The output of the fuzzy controller, as a function of \( x \), will be

  \[
  y_{out} = \sum_{j=1}^{r} m_{A_j}(x)g_j(x) = g(x).
  \]

  \( g(x) \) is the interpolation of the output functions \( g_j(x) \).

  In section 4.5.1 we look at how fuzzy controllers can achieve zero maximum error with a small number of fuzzy if-then rules. This provides insights into how fuzzy controllers’ distinctive feature—the linguistic descriptions of a numerical function—limits the efficiency of fuzzy controllers as function approximators in terms of the number of rules required.

  In section 4.5.2 we consider the case when input fuzzy sets are triangular, output functions are polynomials, and study the rate at which the maximum error goes to zero as the number of fuzzy if-then rules goes to infinity.

### 4.5.1 Achieving Zero Error

Consider the function implemented by the fuzzy controller,

\[
 g(x) = \sum_{j=1}^{r} m_{A_j}(x)g_j(x).
\]

For \( g(x) \), the input membership functions \( \{m_{A_j}(x)\} \) and the output functions \( \{g_j(x)\} \) play almost identical roles. The difference is that \( \{m_{A_j}(x)\} \) must add up to one over
$j$ and take on values between zero and one, but there is no restriction on $\{g_j(x)\}$. Assume that we know $f(x)$, the function to be approximated. (This situation can be approached with more and more training samples.) We observe that:

- **If $g_j(x)$ can be chosen arbitrarily, zero error can be achieved with 1 fuzzy set when $m_{A_1}(x)$ is a constant.**
  
  We can choose $g_1(x) = f(x)$, then $m_{A_1}(x) = 1$, $g(x) = f(x)$, and the error is zero.

- **If $m_{A_j}(x)$ can be chosen arbitrarily, zero error can be achieved with 2 fuzzy sets with $g_j(x) = \mu_j$ for some set of constants $\mu_j$.**
  
  Here we need to make sure that $0 \leq m_{A_j}(x) \leq 1$ and $\sum_j m_{A_j}(x) = 1$. Assume that $f(x)$ takes on the maximum value $f_{\text{max}}$ and the minimum value $f_{\text{min}}$ in the interval of interest, $U = [0, 1]$. We can choose $g_1(x) = f_{\text{min}}$, $g_2(x) = f_{\text{max}}$, and

  $$m_{A_1}(x) = \frac{f_{\text{max}} - f(x)}{f_{\text{max}} - f_{\text{min}}}$$

  $$m_{A_2}(x) = 1 - m_{A_1}(x)$$

  $m_{A_1}(x)$ and $m_{A_2}(x)$ both take on values between 0 and 1 in $x \in [0, 1]$. $g(x) = f(x)$, and the max error is zero.

We learn that:

- **If either the output functions or the input membership functions can be arbitrarily chosen, zero error can be achieved with only one or two rules.**

- **Although zero error can be achieved by either choosing the input membership functions or the output functions in particular ways, doing so sacrifices the distinctive feature of fuzzy controllers: the simple local linguistic descriptions of the numerical function.**

In order for the linguistic fuzzy if-then rules to remain meaningful, easy to be generated, and easy to understand, the input fuzzy sets and the output functions
should remain simple. The input fuzzy membership functions should have similar shapes, usually scaled/shifted versions of each other. This poses restrictions on how much the input fuzzy sets can be customized to fit $f(x)$. The output functions $\{g_j(x)\}$ should be constants if the rules are to remain completely linguistic, so that linguistic expert knowledge can be used directly. This is the case for pure fuzzy controllers.

Given the restrictions that the input fuzzy membership functions and the output functions should be simple, the only way to reduce error is to increase the number of fuzzy if-then rules. In the next section we look at the rate at which maximum error goes to zero as the number of rules increases.

4.5.2 The Convergence Rate for Fuzzy Controllers

In this section our attention is restricted to fuzzy controllers with triangular input fuzzy sets, perhaps the most popular choice of fuzzy membership functions. We assume that:

1. The function approximation takes place on the unit interval, $U = [0, 1]$.

2. The fuzzy controller has $r$ rules:
   
   If $x$ is $A_j$ then $y$ is $g_j(x)$, $j = 1 \ldots r$.

3. The $r$ input fuzzy sets are evenly spaced, triangular-shaped and form a fuzzy partition. Each fuzzy set overlaps with its nearest neighbors only. (see Fig. 4.8)

4. The output functions $\{g_j(x)\}$ are all $k_o$th order polynomials. The most common Sugeno controllers have linear output functions, $k_o = 1$.

For each given $k_o$, the maximum error goes to zero as $r$ goes to infinity. The convergence rate of $g(x)$ to $f(x)$ is defined as the rate at which the maximum error approaches zero as $r$ increases.

---

4It has been argued that Sugeno fuzzy controllers (with $k_o = 1$) require a much smaller number of rules $r$ than pure fuzzy controllers ($k_o = 0$). The convergence rate analysis will show this quantitatively.
Figure 4-8: Fuzzy sets with triangular membership functions.

Suppose fuzzy set $A_j$ peaks at $x = a_j$, $j = 1 \ldots r$. $a_1 = 0$, $a_r = 1$, $a_{j+1} - a_j = 1/(r-1)$. See Fig. 4.8. The function implemented by this fuzzy controller will be the interpolation of two output functions at any point:

$$g(x) = m_{A_j}(x)g_j(x) + m_{A_{j+1}}(x)g_{j+1}(x)$$

$$= \gamma_j(x), \quad x \in [a_j, a_{j+1}], \quad i = 1 \ldots r - 1.$$  

Between the peaks of two neighboring input fuzzy sets, $x \in [a_j, a_{j+1}]$, $g(x)$ is a convex combination of $g_j(x)$, $g_{j+1}(x)$ (i.e. $g(x)$ is the weighted sum of $g_j(x)$ and $g_{j+1}(x)$ where the weighting coefficients add up to one, or $g(x)$ is the interpolation of $g_j(x)$ and $g_{j+1}(x)$). We will call $[a_j, a_{j+1}]$ an interpolation interval. The unit interval is partitioned into $r - 1$ interpolation intervals.

Since $m_{A_j}(x)$ and $m_{A_{j+1}}(x)$ are linear in $[a_j, a_{j+1}]$, $g(x)$ is a $k_0 + 1$th polynomial $\gamma_j(x)$ in the $j$th interpolation interval. Let $k = k_0 + 1$. The fuzzy controller literally divides $[0,1]$ into $r - 1$ equal-sized interpolation intervals, and approximates $f(x)$ in each interval by a $k$th order polynomial, $\gamma_j(x)$.  

When the output functions $\{g_j(x)\}$ are constants, we have a pure fuzzy controller. In this case $k_0 = 0$, $k = 1$. $g(x)$ is a piecewise linear approximation of $f(x)$, see Fig. 4.9 (a). When $\{g_j(x)\}$ are linear, as in the most common Sugeno fuzzy controllers, $g(x)$ is a piecewise second-order polynomial (see Fig. 4.9 (b)).

$^5$There are enough parameters we can choose in $\{g_j(x)\}$ to make $\gamma_j(x)$ any $k$th order polynomial, with the restriction that $\gamma_j(a_{j+1}) = \gamma_{j+1}(a_{j+1})$ (so that $g(x)$ is continuous).
Sugeno fuzzy controllers with linear output functions.

The convergence rate analysis shows that if derivatives of all orders exist, then the maximum error is often best reduced by increasing \( k \) rather than \( r \). This suggests that we use higher order output functions (larger \( k_o \)), or use input fuzzy sets with higher order membership functions (larger \( k_i \)), since \( k = k_i + k_o \). However, one might prefer keeping \( k_i, k_o \) small to preserve the simple linguistic descriptions.

Having linguistic descriptions for the numerical function is the distinctive strength of fuzzy controllers. It also limits the fuzzy controllers as function approximators, since input fuzzy sets and output functions need to be simple. The "performance" should not be measured by the efficiency in minimizing error alone. Ease of expression is an advantage that often outweighs efficiency and precision. If one does not care about the linguistic aspect, then it would indeed be better to use higher order input fuzzy membership functions or output functions. In fact there is no reason to use the structure of fuzzy controllers anymore. Other function approximation schemes may be considered (fourier series, spline, wavelets, neural networks,...).
Chapter 5

Fuzzy Pattern Recognition

5.1 Introduction

In this chapter we apply probabilistic interpretations to fuzzy pattern recognition, with emphasis on fuzzy classification. For more background on fuzzy pattern recognition see [PaM 86] [BeP 92].

Fuzzy sets bring to pattern recognition the ideas of soft decisions and flexible boundaries. We show that these ideas are captured by Bayesian detection. By interpreting fuzzy sets probabilistically, fuzzy pattern recognition can be carried out as Bayesian detection. The benefit of doing so is that prior information can be utilized to better exploit contextual dependency.

5.1.1 The Basic Problem

The basic problem of pattern recognition is: Given an observation $x$ in some pattern space, usually $R^p$, classify it into one of a finite number of classes $C_1, C_2, \ldots, C_k$.

A common way to do the pattern recognition is to define the decision functions $\{f_j\}$. $f_j : R^p \rightarrow R$ will be a function that measures the similarity of $x$ to the class $C_j$, $j = 1, \ldots k$.

When the decision is in the form "$x \in C_j$", it is said to be a hard decision. In this case the pattern space is partitioned into crisp sets, one for each class. When
the decision is represented as the vector of decision function values \( \{ f_j(x) \} \) instead, we have a soft decision. A soft decision can be hardened by deciding that \( x \) is in the class for which the decision function value is the highest.

In fuzzy pattern recognition, classes are modeled as fuzzy sets, and the decision functions are membership functions. Whether the final decision is soft or hard depends on the application. In this chapter we address the case in which the final decision for the fuzzy pattern recognition scheme is hard.

### 5.1.2 Chapter Outline

In section 5.2 the basics of fuzzy pattern recognition are introduced. An example in text recognition reveals that contextual information cannot be fully exploited in conventional fuzzy pattern recognition.

In section 5.3 Bayesian detection is introduced using soft decoding as an example. Fuzzy pattern recognition based on Bayesian detection is proposed. The differences between conventional fuzzy and probabilistic approaches are discussed. Interpreting fuzzy sets probabilistically allows us to combine the best aspects of both approaches.

In section 5.4, we introduce Akra's [Akr 93] automated text recognition scheme. Recognition is viewed as searching for known patterns among the observations. Based on this, Akra used an asymmetric distance as a new measure of similarity. We discuss possible ways to introduce soft decisions (flexible boundaries) into his scheme through fuzzy pattern recognition based on Bayesian hypothesis testing.

### 5.2 Basics of Fuzzy Pattern Recognition

In conventional pattern recognition, each observation \( x \) in the pattern space is classified into exactly one class. In fuzzy pattern recognition, the classes \( \{ C_1, \ldots, C_k \} \) are modeled as fuzzy sets in the pattern space (i.e. the observation space, the space of possible observations), usually assumed to form a fuzzy partition. The observation is classified into multiple classes, each to some degree, indicated by the membership vector \( \{ m_{C_i}(x) \ldots m_{C_k}(x) \} \). The membership vector is the vector of membership
function values of all fuzzy sets at one observation; its components add up to one if the fuzzy sets form a description space (form a fuzzy partition).

The membership vector can be viewed as a fuzzy soft decision. There are other types of soft decisions, e.g. the probabilistic soft decision in soft decoding (see Sec. 5.3).

**Fuzzy Soft Decision:** Membership vector, \{m_{C_1}(X), \ldots m_{C_4}(X)\}.

**Hard Decision:** Most likely class, \(C^* = \arg\max_j m_{C_j}(X)\).

In applications such as remote sensing image pattern recognition [KeM 88], pixels in a satellite image are classified into land cover types. Each pixel corresponds to a large ground area that may contain a mixture of land cover types (e.g. 30 percent bare soil, 70 percent forest). Compared to one-pixel-one-class, it is more natural to give, as the classification result, a fuzzy membership vector which can account for the class mixture.

In most applications of fuzzy pattern recognition, however, the fuzzy decision is hardened to provide a most likely class as the pattern recognition result\(^1\). For example, in text or speech recognition the hard decision is usually more natural; it makes sense to classify the observation into a most likely intended character.

Once the final decision is hardened, the problem is reduced back to conventional pattern recognition. Do we gain from performing fuzzy pattern recognition and then hardening the decision? We can. The gain comes when multiple observations are combined and classified all at once to exploit their contextual dependency. Since soft decisions contain more information, the use of soft intermediate decisions (as in fuzzy recognition) offers advantages over the use of hard intermediate decisions.

Using soft intermediate decisions, as in fuzzy recognition, is better than using hard intermediate decisions since soft decisions contain more information.

Consider the example of fuzzy text recognition [SiC 74][CHC 89]. A character is usually defined as a fuzzy set obtained by combining component fuzzy sets using fuzzy AND and OR.

---

\(^1\)Hardening a fuzzy decision is the dual to "defuzzification". In defuzzification we find a representative observation given the pseudomembership function of a fuzzy set. In hardening a fuzzy decision we find a representative description (fuzzy set) given the membership vector of an observation.
• Fuzzy OR usually accounts for different fonts of the same character. For example, a conceptual character "A" can be viewed as "A, Font 1" OR "A, Font 2" OR ... "A, Font n". Each font of "A" is associated with (but not equal to) a template, one particular pattern in the pattern space, and is modeled as a fuzzy set in the pattern space.

• Fuzzy AND accounts for different strokes (features, primitives) within the same character. "A, Font 1" can be viewed as "stroke 1" AND "stroke 2" AND ... "stroke m", where each conceptual stroke is associated with (but not equal to) a stroke in the template, and is modeled as a fuzzy set over possible observed strokes.

The fuzzy membership function for each stroke is defined in a heuristic way that measures the similarity between the observed stroke and the the candidate stroke in the template. Given an observed character \( x \), the recognition is performed roughly as follows:

1. For each stroke in \( x \), find the stroke membership function value in each candidate stroke.

2. Combine the membership values for the strokes in each candidate character \( C_j \) using fuzzy AND, OR to obtain the overall fuzzy membership function value \( m_{C_j}(x) \).

3. Choose the candidate character with the largest membership function value.

For each observed stroke in an observed character, a soft decision (a fuzzy membership vector) is found that contains more information than a hard decision (the most likely candidate stroke). Combining the component soft decisions within the observed character produces an overall fuzzy soft decision (membership vector) over candidate characters, which is hardened to provide the recognition result.

To classify one object, often several observations are used; they may come from different aspects of the same object (e.g. different strokes in the same object), or from other neighboring objects. Observations of neighboring objects provide information
due to contextual dependency between an object and its neighbors. Exploring the contextual dependency between neighboring observations can improve the classification accuracy. The result will be further improved if soft decisions are used to preserve as much information as possible.

In text recognition, one can combine strokes into characters, characters into words, and so on. In most fuzzy text recognition schemes, however, each character is recognized individually by combining observations over strokes in the same observed character without using observations from any other characters [SiC 74][CHC 89]. More will be gained if observations for different characters are combined. As shown in the example below, conventional fuzzy recognition is unable to fully explore the contextual information.

**Example:** Suppose we observe a sequence of three hand written observations $x$, $y$, $z$, and find that:

1) $x$ looks exactly like $T$: $m_T(x) = 1$.
2) $y$ looks slightly more like $A$ than $H$: $m_A(y) = .51$, $m_H(y) = .49$.
3) $z$ looks exactly like $E$: $m_E(z) = 1$.

If our beliefs about the three observations are combined by “AND” as in conventional fuzzy pattern recognition, we would conclude that

$$m_{TAE}(x, y, z) = \min(m_T(x), m_A(y), m_E(z)) = .51,$$

and

$$m_{THE}(x, y, z) = \min(m_T(x), m_H(y), m_E(z)) = .49.$$

We will decide that $y$ is still slightly more like $A$ than $H$ after using information from $x$ and $z$. The decision for the whole words is completely determined by the decisions for the individual characters.

The component decisions do not account for new information that arises only when all three characters are viewed as a whole: From experience we know that $THE$ occurs much more often than $TAE$ (used in “Tae Kwon Do”). It is natural
to use this prior knowledge to weight the similarities in favor of \textit{THE}. While it is possible to derive heuristic rules to appropriately weigh the combined fuzzy decision, this has already been done systematically within the framework of Bayesian detection, another soft decision approach.

5.3 Bayesian Detection

In Bayesian detection\footnote{We are referring to the most common Bayesian detection scheme, where the decision is to select the candidate class with the maximum aposteriori probability.}, the decision is based on both the similarity of the observation to each candidate class (measured by the likelihood probability) and the prior probability of the candidate class. An example of Bayesian detection is probabilistic soft decoding in communications.

5.4 Probabilistic Soft Decoding as Bayesian Detection

In soft decoding, an observation is viewed as a signal (the transmitted symbol) plus noise. Given a sequence of observations, the most likely sequence of transmitted symbols (a codeword sequence) is selected as follows:

1. For each observation (a received symbol at time \(i, x_i\)), generate \(\{P(x_i|C_j^i)\}\), a vector of likelihood probabilities over possible transmitted symbols \(C_j^i\). This can be interpreted as a probabilistic soft decision.

2. Assuming that the noise variables added to different symbols are independent, the likelihood probability of a sequence of observations will be the product of the individual likelihood probabilities. We can obtain the vector of likelihood probabilities for a sequence of observations, \(P(x_1, \ldots, x_n|C_{j_1}^1 \ldots C_{j_n}^n) = \prod_{i=1}^n P(x_i|C_{j_i}^i)\).
3. In principle, the next step is to weight each term by the prior probability of each codeword sequence (a sequence of possible transmitted symbols) \( P(C_{j_1}^1, \ldots C_{j_n}^n) \)

The vector \( \{P(x_1, \ldots x_n | C_{j_1}^1 \ldots C_{j_n}^n) P(C_{j_1}^1, \ldots C_{j_n}^n)\} \) can be hardened to give, as a final decision, the codeword sequence with the largest aposteriori probability. If this vector is normalized (by multiplying each term by a constant) so that all entries add up to one, the result is an aposteriori probability vector \( \{P(C_{j_1}^1 \ldots C_{j_n}^n | x_1 \ldots x_n)\} \). The selected codeword sequence will be the same whether or not the normalization is done.

The maximum aposteriori probability (MAP) decision described above minimizes the Bayes risk, defined as the probability of error in this case.

Fuzzy pattern recognition is similar to soft decoding. Both give soft component decisions (membership vector or vector of likelihood probabilities), combine several of them, then harden the decision at the end. Often the fuzzy sets are assumed to form a fuzzy partition,

\[
\sum_j m_{C_j}(x) = 1;
\]

in this case the membership vector can be interpreted as an aposteriori probability vector:

\[ m_{C_j}(x) = P(C_j | x). \]

The fuzzy sets can be viewed as the underlying hypotheses that generate the observations.

The main differences between Bayesian detection and conventional fuzzy pattern recognition are:

1. **The component decisions and the way they are combined.** In conventional fuzzy recognition, component membership functions are usually combined by fuzzy AND, OR, which may correspond to \( \min/\max, \text{product/sum minus product} \), and so on. In Bayesian detection, component likelihood probabilities

\[ ^3 \text{Usually all codeword sequences have the same nonzero prior probabilities, and all sequences not in the codebook have zero prior probability. Therefore one can find the likelihood probabilities for codeword sequences and use them directly for making the decision.} \]
(which can be interpreted as pseudomembership functions) are usually combined by product.

2. The weighting by prior probabilities. In Bayesian detection, the combined decision (likelihood probability of the sequence of observations given the sequence of hypotheses) is weighted by the prior probability of the sequence of hypotheses (e.g. the codeword sequence in soft decoding) to obtain the maximum aposteriori probability (MAP) decision. There is no corresponding weighting in fuzzy recognition. Weighting by prior probabilities is important for fully exploring contextual information, as seen in the previous text recognition example.

5.4.1 Fuzzy Pattern Recognition as Bayesian Detection

The probabilistic interpretations of fuzzy sets allow us to perform fuzzy pattern recognition as Bayesian detection, so that contextual information can be better utilized. Each candidate class will be modeled as a detection-type fuzzy set (a hypothesis), based on which the observation is generated. The recognition scheme still selects the fuzzy set with the largest membership function value. However, the membership function is specified indirectly through the pseudomembership function and the prior probabilities by means of Bayes rule. The correspondence between fuzzy recognition and Bayesian detection are:

- Fuzzy set (detection-type) $\leftrightarrow$ Hypothesis.
- Pseudomembership functions $\leftrightarrow$ Likelihood probabilities.
- Membership functions $\leftrightarrow$ A posteriori probabilities.
- Select fuzzy set with largest membership function value $\leftrightarrow$ Select hypothesis with maximum a posteriori probability.

Fuzzy pattern recognition can be carried out as follows:
1. Similarities of the observation to the candidate classes will first be measured in terms of pseudomembership functions (as opposed to membership functions), which correspond to the likelihood probabilities in Bayesian detection.

2. Assume that an observation, given the underlying fuzzy set, is conditionally independent of the other observations and their underlying fuzzy sets (see Section 3.6.2). Combine the pseudomembership function values (likelihood probabilities) of several observations by taking the product to give the pseudomembership function value of the intersection of component fuzzy sets at the observations. This corresponds to the likelihood probability of the sequence of observations given the sequence of hypotheses.

3. Each resultant pseudomembership function (likelihood probability) is weighted by the prior probability of the sequence of hypotheses to make the decision that maximizes the membership function value (aposteriori probability).

When more observations are combined, it is easy operationally to find the pseudomembership functions, and then weight by the prior probabilities at the top level. However, as the number of observations increase, the possible combinations of underlying fuzzy sets will grow exponentially. We may deal with this by going back to the membership functions and eliminating unlikely choices at each level 4.

**Fuzzy and Probabilistic Approaches**

When fuzzy pattern recognition is modified based on Bayesian detection, it reduces to the probabilistic approach computationally. Fuzzy and random ideas are conceptually different, however. Within the framework of fuzzy recognition, the generation of an observation given an underlying hypothesis is no longer viewed as “template plus noise”, but as “choosing a style”.

---

4In soft decoding, the number of candidate codeword sequences also grows exponentially as the number of observations grows. In implementation, the Viterbi algorithm trims down the number of candidate codeword sequences, keeping only a fixed number of leading candidates as new observations come in.
In pattern recognition, each class usually has a template which is a typical or ideal observation for that class (e.g. the average vector of that class in the pattern space). In a simple minded probabilistic approach, the observation is often viewed as the sum of a template and some independent additive noise, e.g. \( X = T_A + N \) (signal plus noise). In a more sophisticated model, the observation is obtained by deforming the template (deformation can be template dependent), then adding independent noise. In all of these cases, the template is always where one starts. The likelihood probability is \( P(x|A) = P(x|T_A) \). The event "A occurs" corresponds to "\( T_A \) occurs".

In fuzzy pattern recognition, the conceptual character \( A \) is viewed as a fuzzy set. Although the template \( T_A \) can be found by defuzzifying the pseudomembership function, \( A \) does not correspond to one single pattern, the template \( T_A \). Instead, \( A \) is viewed as a fuzzy set that contains many realizations, each more or less like \( A \). When modeled as an event in the enlarged sample space, "A occurs" does not corresponds to any particular written pattern, but simply leads to "the plausibility that \( x \) will be observed is \( pm_A(x) = P(x|A) \)". Conceptually, there is no need to view \( x \) as being obtained from applying deformation and noise to one particular template \( T_A \). It is more natural to view the varieties of \( x \) as the characteristic of \( A \), and the generation of \( x \) as choosing a style and particular representation based on \( pm_A(x) = P(x|A) \). The pseudomembership function \( pm_A(x) \) provides a numerical measure of the plausibility of \( x \) appearing under the hypothesis \( A \).

Although the idea of a template is not conceptually essential, it can be helpful in choosing pseudomembership functions. Pseudomembership functions can be defined directly based on parameters that measure similarities between the template and the observation. Viewing the observation as template perturbed by various types of noise can help one gain intuition, but can lead to unnecessary complexity.

**Combining the Best of Fuzzy and Probabilistic Approaches**

Note that in casting fuzzy pattern recognition within a probabilistic framework, we are not advocating using "real" probabilities that are found through a probabilistic model that is constructed to account for the way observations are physically generated. Such
a probabilistic model can contain too much detail and, although helpful conceptually, is not too useful since the probabilities can be hard to find (e.g. see the example in the last section).

In fuzzy recognition, one instead first identifies important parameters that are obtainable and measures the similarities between observations and the candidate classes. One then defines fuzzy sets based on these parameters. It is not necessary to describe the way observations are physically generated.

What we suggest is to combine the advantages of both fuzzy and probabilistic approaches. Having a conceptual probabilistic model and performing fuzzy recognition as Bayesian detection have these advantages:

- With the probabilistic approach, both the similarities between the observations and the candidate classes (the likelihood probabilities) and our knowledge about the candidate classes from past experience (the prior probabilities)\(^5\) will be used for decision making, in order to fully explore contextual information and increase the recognition accuracy over conventional fuzzy recognition.

- With the fuzzy approach, the measure of similarity will be defined in a convenient way that agrees with our engineering insight. The task of finding the likelihood probabilities, which can be tedious in a conventional probabilistic approach, is greatly simplified.

5.5 Example: Automated Text Recognition

5.5.1 Akra’s Scheme

Akra[Akr 93] proposed a way to perform automated recognition of handwritten text based on the following principles.

The Layering Principle: The recognition process is a layered hierarchical process

---

\(^5\)Note that the prior probability does not depend on the current observation, although it can be updated based on the classification result.
representing different levels of abstraction. Sets of points are recognized as characters, sets of characters as words, and sets of words as concepts.

This is an established principle in the text recognition field. In addition to the layering principle, Akra proposed several other principles applicable to each level of the hierarchy exhibited by the layering principle.

The Searching Principle: The recognition process is not a complete matching between what we know and what we observe. Rather, it is a search for what we know in what we observe.

Recognition is asymmetric. We look for our known object (template) in the environment (sample/observation), ignoring the part that doesn’t make sense. The “clean” knowledge (template) is more important than the “noisy” observations (data). This suggests that a metric is not a good choice for measuring similarity, as it is symmetric between the template and the observation. Akra [Akr 93] instead proposed the single sided Hausdorff distance from the template to the observation as the measure of similarity.

Definition: Let $T = \{ \alpha_1, \alpha_2, \ldots \alpha_M \}$ and $S = \{ \beta_1, \beta_2, \ldots \beta_N \}$ be two finite sets of points in $R^2$. Let $d$ be the Euclidian metric in $R^2$. Let $h(T, S)$ be calculated as follows:

$$h(T, S) = \max_{i} \min_{j} d(\alpha_i, \beta_j)$$

Then $h(T, S)$ is the single sided Hausdorff distance from $T$ to $S$.

When $T$ is the template of a conceptual character, and $S$ is the observed sample, the single sided Hausdorff distance $h(T, S)$ measures how much of the template can be found in the observation, ignoring everything extra. The conventional Hausdorff distance is defined as $H(T, S) = \max(h(T, S), h(S, T))$. Note that $h(T, S) = 0 \iff T \subset S$ ($T$ can be found in $S$), whereas $H(T, S) = 0 \iff T = S$. Based on the search

116
principle, the single sided Hausdorff distance is proposed as a superior measure of similarity compared to the Hausdorff distance.

![Diagram of H's with different observations]

Figure 5-1: Recognition as Searching for Patterns.

Examples illustrating the searching principle are shown in Fig. 5.1. The single sided Hausdorff distance is zero for all five observations. The template for $H$ (without serifs) can be found in an $H$ written with serifs (observation 1), with blobs (observation 2), with scratches (observation 3), or touching another character (observation 4). Observation 5 contains the template of $H$, but also looks like an $A$. Akra [Akr 93] proposed another principle for this situation.

The Maximum Understanding Principle: Recognition is a process of searching for the known pattern that accounts for the maximum amount of the observation.

In observation 5, the maximum understanding principle favors an interpretation of $A$ over $H$ since the first interpretation accounts for every stroke in the observation. The same principle applies to the word level. For example, consider the interpretation of the word “smile”, which contains the words “mile” and “smile”. We would choose “smile” since it offers the maximum understanding possible from the observation.
The maximum understanding principle in a sense compensates for the use of single sided Hausdorff distance, which ignores all extra points including ones that might make sense.

The final principle of recognition Akra proposed is:

**The Allowable Deformation Principle:** We recognize a shape if we can reach it by deforming a template using some chain of possibly template-dependent rules of deformation within the allowable range.

A character shape $A$ (the template) emerges from a collection of points (the observation) if there are certain allowable deformations (e.g., slanting by 30 degrees, disturbance of points by 1 percent of the character size, etc.) that make the character $A$ match a subset of the collection of points (the observation).

Within the subsequent discussions, a character will be represented as follows: The image of a character is assumed to be black and white. It is quantized into a finite number of pixels, each represented by a two dimensional coordinate vector. The vector of coordinate vectors of the pixels that are black corresponds to a character $T = \{\alpha_1, \ldots, \alpha_M\}$.

**Definition:** Let $T = \{\alpha_1, \ldots, \alpha_M\}$ where $\alpha_i \in R^2$, for $1 \leq i \leq M$. A translation of $T$ by $t \in R^2$ is denoted $T \oplus t$ and is defined as the set:

$$\{\alpha_1 + t, \alpha_2 + t, \ldots, \alpha_M + t\}$$

**Definition:** Let $T$ be a template. A deformation $P$ of $T$ is called linear if it can be represented as premultiplication of each coordinate vector in $T$ by a fixed $2 \times 2$ matrix $P$.

The linear deformations considered in [Akr 93] include: Scaling $(P = \begin{pmatrix} \alpha_x & 0 \\ 0 & \alpha_x \end{pmatrix})$, slanting $(P = \begin{pmatrix} 1 & \tan \alpha_s \\ 0 & 1 \end{pmatrix})$, and rotation $(P = \begin{pmatrix} \cos \alpha_r & \sin \alpha_r \\ -\sin \alpha_r & \cos \alpha_r \end{pmatrix})$.

The scaling parameters $\alpha_x, \alpha_y, \alpha_s, \alpha_r$ are each restricted to some allowable in-
tervals that might be template dependent.

Let $X_T$ be the space of allowable linear deformations corresponding to a template $T$. The problem of recognition then becomes finding a path (a sequence of linear deformations) in $X_T$ that, with appropriate translation, matches $T$ to a subset of the observation $S$. A template $T$ emerges from an observation $S$ if the reduced single sided Hausdorff distance \(^6\), $h^*(T, S)$ is less than a constant:

$$h^*(T, S) = \min_{P \in X_T} \min_t h(PT \oplus t, S) < \epsilon.$$  

For complexity reasons [Akr 93] considers only the above character deformations before taking the single sided Hausdorff distance. Conceptually word deformation, local deformation (on each stroke) and amplitude deformation can also be considered. [Akr 93]'s algorithm accepts as input an observation $S$. To recognize a template $T$, the algorithm translates, scales, slants, and rotates $T$ in order to reduce the single sided Hausdorff distance (within the allowable range) until reduction in the distance is no longer possible. If the resultant single sided Hausdorff distance is less than some constant, the character is recognized, otherwise it is not. That is, the character $S$ is recognized only if it can be "reached" from $T$ by applying the deformations mentioned above to reduce the single sided Hausdorff measure to a quantity no bigger than $\epsilon$.

5.5.2 Adopting Soft Decisions

In Akra’s scheme, the parameter used for recognition is the reduced single sided Hausdorff distance from the appropriately deformed character template to the observation. A conceptual character $C_j$ is either recognized or not recognized; it is recognized if the reduced single sided Hausdorff distance is within a crisp interval $[0, \epsilon_{C_j}]$. An infinitesimal variation at the boundaries can make the difference of recognition or not.

\(^6\)The conceptual idea of text recognition based on a reduced distance (after optimal deformations are applied) was mentioned in [BKZ 66]. The expression of single sided Hausdorff distance appeared in [BKV 66], but the authors did not recognize that this is not a symmetrical distance. Recognition is still viewed as the matching between two patterns.
It is more natural to introduce gradual boundaries through fuzzy sets.

In conventional fuzzy pattern recognition, one would define a fuzzy membership function directly to replace a crisp interval. In fuzzy recognition based on Bayesian detection, this would be done indirectly, starting by defining the pseudomembership function. Akra's threshold $c_{ij}$ for class $C_j$ is replaced by defining the pseudomembership function $pm_{C_j}(x)$ (likelihood probability) for each conceptual character $C_j$. Together with the prior probabilities, the membership functions can be found using Bayes rule.

Defining and evaluating the pseudomembership functions (likelihood probabilities) based on a conventional probabilistic model can be difficult (see appendix B). We will instead follow Akra by defining a pseudomembership function based on the single sided Hausdorff distance. The next example illustrates a weakness of single sided Hausdorff measure, and a heuristic way to compensate for it.

**Example:** Due to the nature of the single sided Hausdorff distance, characters that are simple have an advantage of being recognized (being found in the observation) over characters that are more complicated. Consider the example in Fig. 5.2. $h^*(T_I, x) = 0 < h^*(T_H, x)$, even though the observation clearly looks more like $H$.

![Figure 5-2: The reduced single sided Hausdorff distance is not perfect.](image)

Akra dealt with this through the choice of different thresholds for the reduced single sided Hausdorff distance below which recognition occurs, and choosing the most complicated character when multiple characters are recognized.

In our view, the situation in Fig. 5.2 can be accounted for if the whole vector of reduced single sided Hausdorff distance is considered in finding the similarity between
the observation and each candidate character. Assume for simplicity that there are only two candidate characters, I and H. Then the similarities can be measured by the probabilities:

\[ P(h^*(T_I, x), h^*(T_H, x)|I), \]

\[ P(h^*(T_I, x), h^*(T_H, x)|H). \]

Given the observation \( x \) in Fig. 5.2, \( P(h^*(T_I, x), h^*(T_H, x)|I) \) would be small even though \( h^*(T_I, x) \) is small, since \( h^*(T_H, x) \) is also small, which is unlikely if \( x \) comes from \( I \).

When there are \( k \) candidate characters (\( k \) would be at least 36 counting letters plus numerals), it can be too complicated to find \( P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x)|C_j) \) as an indication of the similarity of \( x \) to \( C_j \). An alternative is to approximate \( P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x)|C_j) \) with a function of \( h^*(T_{C_j}, x) \) only.

To compensate for the bias against complicated characters, the complexity of character \( C_j \) can be used in defining the function that approximates \( P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x)|C_j) \). For example, let \( n(T_{C_j}) \) be the number of black pixels in \( T_{C_j} \). Thus \( n(T_{C_j}) \) is a measure of complexity of \( T_{C_j} \). We can define the normalized reduced single sided Hausdorff distance as:

\[ h^*_n(T_{C_j}, x) = \frac{h^*(T_{C_j}, x)|C_j}{n(T_{C_j})} + \frac{c}{n(T_{C_j})} = \frac{h^*(T_{C_j}, x)|C_j + c}{n(T_{C_j})}, \]

where \( c \) is a positive constant. With a proper choice of \( c \), the observation \( x \) in Fig. 5.2 will have \( h^*_n(T_I, x) > h^*_n(T_H, x) \). Note that the above is one of many possible ways to normalize the reduced single sided Hausdorff distance.

We can then define some simple decreasing function \( f(h^*_n(T_{C_j}, x)) \) to approximate

\footnote{We are dealing with probabilities, not probability densities, since the numbers of possible \( x \) and \( h^*(T_{C_j}, x) \) are finite (the total number of pixels in the region where the character is written is finite).}
the likelihood probability of the vector of reduced single sided Hausdorff distances:

\[ P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x) | C_j) \simeq constant * f(h^*_n(T_{C_j}, x)). \]

In the Bayesian detection, \( pm_{C_j}(x) = P(x | C_j) \) can be replaced by \( f(h^*_n(T_{C_j}, x)) \) based on the belief that

\[ pm_{C_j}(x) = P(x | C_j) \simeq constant * f(h^*_n(T_{C_j}, x)). \]

Multiplying by a constant does not change the relatively likelihood among classes, and it is unnecessary to find the value of the constant. Note that:

- We assume that the vector of reduced single sided Hausdorff distance \( \{h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x)\} \) captures the similarity of \( x \) to \( C_j \). In other words,

\[ pm_{C_j}(x) = p(x | C_j) \simeq constant * P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x) | C_j). \]

The constant accounts for the fact that there are multiple \( x \)'s with the same value of \( \{h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x)\} \). In a sense, \( x \)'s are "quantized" according to \( \{h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x)\} \).

- It can be difficult to compute \( P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x) | C_j) \). We further approximate it with \( f(h^*_n(T_{C_j}, x)) \), which means

\[ P(h^*(T_{C_1}, x), \ldots, h^*(T_{C_k}, x) | C_j) \simeq constant * f(h^*_n(T_{C_j}, x)). \]

The definition and approximation of the pseudomembership function is rather subtle. There are many questions left. It is hard to make precise the relation \( P(x | C_j) \simeq constant * f(h^*_n(T_{C_j}, x)) \). Further research is clearly required, perhaps through experiments.

Assuming that we can obtain the pseudomembership functions or their approximations (multiplied by a constant), the pseudomembership function (likelihood probability) given each class will be multiplied by the prior probability, \( P(C_j) \). The decision
vector is \( \{P(x|C_j)P(C_j)\} \); this can be normalized so that the components add up to one, in which case it becomes a membership vector \( \{P(C_j|x)\} \). Normalization is not necessary if we only want the class with the largest membership function value.

To use contextual information, we can combine the likelihood vectors for multiple observations and weight each term by the joint prior probability on the top level. For example, in combining \( x_1, \ldots x_k \) to recognize \( C_{j_1}^1, \ldots C_{j_k}^k \), we would compute the likelihood vector

\[
\{P(x_1 \ldots x_k|C_{j_1}^1 \ldots C_{j_k}^k) = P(x_1|C_{j_1}^1) \ldots P(x_k|C_{j_k}^k)\}.
\]

We could then weight each term by \( P(C_{j_1}^1 \ldots C_{j_k}^k) \) to obtain the decision vector

\[
\{P(x_1 \ldots x_k|C_{j_1}^1 \ldots C_{j_k}^k)P(C_{j_1}^1 \ldots C_{j_k}^k)\},
\]

which can also be normalized to become a membership vector, if required.

As shown in Fig. 5.3, the text recognition can be performed in a hierarchical manner. Characters are grouped into words, words are grouped into sentences, and so on. A vector of likelihood probabilities will be computed for each node. These vectors will be propagated upward, and on the top level the final decision is made after weighting each term in the vector by the prior probability. In a naive implementation, the size of the likelihood vector will grow exponentially as we go up the levels. This can be dealt with by trimming down all but a few leading candidates at each level. The leading candidates are those with highest membership function values. Finding them requires weighting the likelihood vector with prior probabilities at each level.

### 5.6 Conclusions

This chapter proposed a framework for performing fuzzy pattern recognition based on Bayesian detection. Fuzzy pattern recognition further illustrates the advantage of combining the insight in fuzzy set theory and the mathematical structure of probability, which is made possible by the probabilistic interpretation of fuzzy sets.
Figure 5-3: The hierarchical structure of text recognition.

**Contribution of probability: Better use of contextual information.**

Fuzzy pattern recognition can correspond to Bayesian detection. The detection-type fuzzy sets that model candidate classes can be viewed as underlying hypotheses. Pseudomembership functions correspond to likelihood probabilities, and membership functions correspond to a posteriori probabilities. Probability suggests using both the similarity of the observation to the candidate class (likelihood probability) and the prior information about the candidate class (prior probability) in making a decision. The prior probability is obtained from past experience; it depends only on the candidate class but not on the observation. The prior probability is important in exploiting contextual dependency, and will lead to improved recognition accuracy.
Contribution of fuzzy set theory: Simplification of the probability model.

A conventional probabilistic model attempts to describe how observations are physically generated. Finding the likelihood probabilities according to this model can be very difficult. In the fuzzy set approach, membership functions are defined intuitively to measure similarity between patterns based on parameters that can be obtained reasonably. Based on ideas from fuzzy set theory, the likelihood probability in Bayesian detection will be viewed simply as a measure of similarity. The conceptual probabilistic model no longer needs to describe how observations are generated, and the likelihood probability (pseudomembership function) can be defined in an approximate way that agrees with our engineering intuition.

As an example, we studied Akra's text recognition scheme [Akr 93]. In [Akr 93], the similarity of an observation to a candidate class is measured by an asymmetric distance, based on the philosophy that pattern recognition is more like searching for known patterns in what we observe than a complete matching of what we know and what we observe. We suggested a way to introduce soft decisions into Akra's scheme with fuzzy pattern recognition based on Bayesian detection.
Chapter 6

Conclusion

Probability was developed to deal with randomness. Fuzzy set theory was developed to deal with fuzziness (imprecision). Although randomness and fuzziness are uncertainties of different natures, the same mathematical structure—that of probability—can be employed to deal with both fuzziness and randomness.

The purpose of this thesis has been to gain more insight into fuzzy sets and systems using probability as a guide. In the beginning of the thesis, we discussed the difficulties in fuzzy set theory: the validity of statements such as “If A then not A” when A is not empty, and the proliferation of many ad hoc fuzzy union/intersection operations. These issues are resolved when fuzzy sets are interpreted probabilistically. Adopting the mathematical structure of probability strengthens fuzzy set theory.

We proposed a probabilistic interpretation of fuzzy sets, and examined the impact on fuzzy control and fuzzy pattern recognition. Our approach differs from others in that the idea of the enlarged sample space is clarified, which is essential for understanding how fuzziness is modeled probabilistically.

Probabilistic Interpretations of Fuzzy Sets

- Any probabilistic model for fuzzy sets employs what we call an “enlarged sample space”. The enlarged sample space can be viewed as the Cartesian product of an observation space (the collection of observations, the “universe of discourse” in conventional fuzzy set theory) and a description space (the collection of fuzzy
sets including intersections and complements).

- Fuzzy sets can be viewed as events (crisp sets) in the enlarged sample space. The membership functions can be viewed as conditional probabilities.

- Fuzzy union/intersection are taken according to the relations between fuzzy sets, as opposed to using the same pair of fuzzy union/intersection operations for any two fuzzy sets. Fuzzy sets, as events in the enlarged sample space, satisfy all basic rules in set theory, including the law of excluded middle, the law of contradiction, and the law of self contradiction.

- There are two basic types of fuzzy sets: Quantization-type fuzzy sets where membership functions are more natural (e.g. fuzzy set \( A \) with \( m_A(x) = P(A|x) \)), and detection-type fuzzy sets where pseudomembership functions are more natural (e.g. fuzzy set \( B \) with \( pm_B(x) = P(x|B) \)). Defuzzification is performed using pseudomembership functions.

**Probabilistic Interpretations of Fuzzy Control**

Fuzzy control, with its linguistic rules, allows the intuition of experts to be used effectively. Adding a probabilistic interpretation helps to sharpen these insights without removing the linguistic advantages.

- The operations of a fuzzy controller can be viewed conceptually as the occurrence of a sequence of probabilistic events. The defuzzified output corresponds to a conditional mean or conditional mode, depending on the defuzzification scheme.

- Probability suggests using product fuzzification, product inference, and sum composition.

- Probability reveals that input fuzzy sets are naturally modeled as quantization-type fuzzy sets forming a partition on the enlarged sample space (all membership
functions add up to one at each observation). Output fuzzy sets are detection-
type fuzzy sets; they should be described by the pseudomembership functions
(each pseudomembership function adds up to one over all observations).

- Existing fuzzy controllers with product fuzzification/inference, sum com-
composition, and centroid defuzzification always conform to this probabilistic structure,
due to the normalization in defuzzification. It is the fuzzy sets after normaliza-
tion that determine the behavior of the fuzzy controller.

**Probabilistic Interpretations of Fuzzy Pattern Recognition**

- When the desirable decision is the most likely candidate class, fuzzy pattern
recognition can be carried out as Bayesian detection. On the other hand, con-
ventional Bayesian detection can be viewed as fuzzy pattern recognition. The
fuzzy sets are underlying hypotheses, pseudomembership functions are likeli-
hood probabilities, and membership functions are a posteriori probabilities.

- Conventional fuzzy pattern recognition lacks the ability to fully exploit con-
textual dependency. Bayesian detection allows the use of prior probability to
better utilize contextual information.

- Conventional Bayesian detection, where the observation is modeled as “template
plus noise”, can be unnecessarily complex. Likelihood probabilities, which mea-
sure the similarity of the observation to the candidate classes, can be difficult
to compute. Fuzzy pattern recognition suggests measuring similarity using pa-
rameters that are reasonable to obtain and agree with our engineering insight.
Likelihood probabilities (in a conceptual probabilistic model) can be defined in
terms of these parameters in an approximate way, and can sharpen one’s insight
about the chosen parameters.

**Fuzzy Set Theory and Probability: a Unified Treatment for Uncertainty.**

Although the mathematical structure of probability can be employed to deal with
fuzzy sets, this is not to say that one can do without fuzzy set theory and simply
solve all problems using probability. The mathematical structure of probability theory can be used to prevent contradictions and paradoxes in the use of fuzzy sets, but the emphasis on linguistic interpretation and insight in fuzzy set theory can be useful in probability. Integrating the two areas produces a unified framework to deal with uncertainty, including fuzziness and randomness. This framework offers great room for further exploration, particularly in studying the impact of probabilistic interpretations on various other fuzzy set applications and problems that contain both fuzzy and random components.
Appendix A

The Convergence Rate

Suppose \( f(X) \) is a function defined over \( U = [0, 1] \). \( g(X) \) is a piecewise \( k \)th order polynomial approximation of \( f(X) \), obtained in the following way. Divide the unit interval into \( r - 1 \) equal-sized interpolation intervals. In each interpolation interval, there are \( k + 1 \) equally spaced points, including the two boundary points, where \( g(X) \), a \( k \)th order interpolating polynomial of \( f(X) \), is selected to agree with \( f(X) \).

We want to find the maximum error between \( f(x) \) and \( g(x) \) as \( r \) goes to infinity and \( k \) is kept small. Suppose \( f(x) \) and \( g(x) \) agree at \( \{x_0, x_1, \ldots x_n\} \), \( x_{i+1} - x_i = h \), \( x_0 = 0 \), \( x_n = 1 \), \( h = 1/n \). For fixed \( k \) and very large \( r \), \( r = O(n/k) = O(n) \), where \( n + 1 \) is the number of interpolating points. The convergence rate would be the same written in terms of \( r \) or \( n \).

In terms of fuzzy controllers, \( f(x) \) is the ideal control function to be approximated by the fuzzy controller, and \( g(x) \) is the function actually implemented by the fuzzy controller. Also,

- \( r \) is the number of fuzzy if-then rules;
- \( k = k_i + k_o \) is the sum of the order of output functions and input fuzzy membership functions, assuming the output functions to be \( k_o \)th order polynomials, and the input fuzzy membership functions to be piecewise \( k_i \)th order polynomials. When the input fuzzy sets are triangular, \( k_i = 1 \), \( k = k_o + 1 \).
- \( n + 1 \) is the number of points, equally spaced in \( U = [0, 1] \), where \( g(x) \) and \( f(x) \)
agree. There are \( k + 1 \) such points in each of the \( r - 1 \) interpolation interval, including the end points, therefore \( n + 1 = k(r - 1) + 1 \), or \( n = k(r - 1) \). As \( r \) goes to infinity and \( k \) is kept small, \( n = O(r) \).

- \( h \) is the distance between two neighboring points, i.e., points at which \( g(x) \) is selected to agree with \( f(x) \). Thus \( nh = 1 \), or \( h = 1/n = 1/k(r - 1) \).

The following derivation is from section 3.6 in "Applied Numerical Analysis" by Gerald and Wheatley [GeW 84].

We start by analyzing the first interpolation interval, \([0, 1/r] \), \( x_0 = 0 \), and \( x_k = 1/r \). Let \( f(x) \) be a function and \( P_k(x) \) be a \( k \)th-degree interpolating polynomial that matches with \( f(x) \) at \( k + 1 \) evenly spaced points \( \{x_0, \ldots, x_k\} \). \( g(x) = P_k(x), x \in [x_0, x_k] \). \([x_0, x_k] = [0, 1/r] \) is one of the interpolation intervals over which \( g(x) \) is a \( k \)th degree interpolating polynomial of \( f(x) \). Assume that \( f(x) \) has a continuous \( k + 1 \)st order derivative on the interval \([x_0, x_n] = [0, 1] \). We call the error function \( E(x) \), where \( 0 \leq x \leq x_k \).

\[
E(x) = f(x) - P_k(x) = (x - x_0)(x - x_1)\cdots(x - x_k)q(x).
\]

Obviously, \( f(x) - P_k(x) - E(x) = 0 \), so

\[
f(x) - P_k(x) - (x - x_0)(x - x_1)\cdots(x - x_k)q(x) = 0.
\]

To determine \( q(x) \), construct an auxiliary function \( W(t) \):

\[
W(t) = f(t) - P_k(t) - (t - x_0)\cdots(t - x_k)q(x).
\]

Note that \( x \) has not been replaced by \( t \) in the \( q(x) \) portion. \( W(t) \) is a function of both \( t \) and \( x \). It is written as a function of \( t \) for convenience, since in the following we will vary \( t \) while keeping \( x \) constant. We next examine the zeros of \( W(t) \).

Assuming that \( x \) is different from each of the points \( x_0, \ldots, x_k \), there are altogether \( k + 2 \) values of \( t \) that make \( W(t) = 0 \), i.e., \( t = x_0, x_1, \ldots, x_k, x \). We now impose
the necessary requirements on \( W(t) \) for the law of mean value to hold. \( W(t) \) must have a continuous derivative. If this is so, there is a zero to its derivative, \( W'(t) \), between each of the \( k + 2 \) zeros of \( W(t) \), a total of \( k + 1 \) zeros. If \( W''(t) \) exists and is continuous, there will be \( k \) zeros of \( W''(t) \), and likewise \( k - 1 \) zeros of \( W'''(t) \), etc., until we reach \( W^{(k+1)}(t) \), which has at least one zero in the interval that has \( x_0, x_k \) as end points. Let this value of \( t \) be denoted \( \xi \). We then have

\[
W^{(k+1)}(\xi) = 0 = \frac{d^{k+1}}{dt^{k+1}}[f(t) - P_k(t) - (t - x_0) \cdots (t - x_k)q(x)]_{t=\xi} = f^{(k+1)}(\xi) - 0 - (k + 1)! \cdot q(x).
\]

Therefore,

\[
q(x) = \frac{f^{(k+1)}(\xi)}{(k + 1)!}, \quad \xi \text{ between } (x_0, x_k).
\]

We now have our error term:

\[
E(x) = (x - x_0)(x - x_1) \cdots (x - x_k) \frac{f^{(k+1)}(\xi)}{(k + 1)!}.
\]

We can upper bound the maximum error in \([x_0, x_k]\), \( \max_{x \in [x_0, x_k]} |E(x)| \), by noting that:

1) \( |x - x_i| \leq h = 1/n \) for \( i = 0 \ldots k \).

2) There exists a nonnegative, finite constant \( c \) such that \( |f^{(k+1)}(\xi)| \leq c \), for all \( \xi \in [x_0, x_n] = [0, 1] \). This follows from our assumption: \( f(x) \) has a continuous \( k + 1 \)st derivative in \([x_0, x_k]\).

As a result,

\[
\max_{x \in [x_0, x_k]} |E(X)| \leq n^{-(k+1)} \frac{c}{(k + 1)!} = [(r - 1)k]^{-(k+1)} \frac{c}{(k + 1)!}.
\]

g(x) consists of \( r - 1 \) kth order interpolating polynomials \( \{P^1_k(x) \ldots P^{r-1}_k(x)\} \), one for each interpolation interval. The maximum error between \( f(x) \) and \( g(x) \) over \([x_0, x_n]\), \( \max_{x \in [x_0, x_n]} |E(x)| \), will be the maximum over \( r - 1 \) error.
terms as above, and each one is $O(n^{-k+1})$. $k$ is constant and does not vary as $n$ and $r$ go to infinity.

$$\max_{x \in [x_0, x_n]} |E(x)| = \max\{ \max_{j} \max_{x \in [x_{(j-1)k}, x_{jk}]} |E(x)| \}$$

Taking the maximum over the $r - 1$ error terms, the overall maximum error is still $O(n^{-k+1})$. In terms of $r$, the convergence rate is

$$O((k(r - 1))^{-(k+1)}) = O(k^{-(k+1)}(r - 1)^{-(k+1)}).$$

Since $k$ remains fixed as $r$ goes to infinity, $k^{-(k+1)}$ can be viewed as a constant, and the convergence rate can be written as:

$$O(r^{-(k+1)}).$$
Appendix B

A Conventional Probabilistic Model for Text Generation

In a purely probabilistic approach, the generation of the written characters can be modeled as shown in the figure:

![Diagram](image)

Figure B-1: A conventional probabilistic model of the character generation process.

Given a conceptual character (a hypothesis $C_j$), a particular style $y$ of writing $C_j$ is chosen according to $P(y|C_j)$. This can be modeled as applying deformations...
(scaling, rotation, slanting, ...) to $T_j$, the template of $C_j$. Noise is then added to $y$ to generate the observation $X$. There may be two types of noise: Pointwise translation (e.g. due to an unsteady hand or printer head), and turning pixels on/off (e.g. from a photocopy or fax machine).

This probabilistic model is helpful in understanding the character generation process conceptually. It is not very practical for use in the recognition algorithm, though. In order to find the likelihood probability $P_{X|C_j}$, one would need to integrate $P_{X,Y|C_j} = P_{Y|C_j}P_{X|Y,C_j}$ over all possible $y$, which is difficult since $P_{X,Y|C_j}$ is not a closed form function of $x$, $y$. One would need to physically deform $T_j$ to create all possible $y$ in order to find $P_{Y|C_j}P_{X|Y,C_j}$ for all $y$, which is tedious and not realistic. It is much simpler if one takes the fuzzy view and bases the recognition on the normalized reduced single sided Hausdorff distance. The pseudomembership function defined based on this parameter can be viewed as the likelihood probability in a conceptual probabilistic experiment. The pseudomembership function $pm_{C_j}(x)$ may not correspond to the physically measurable frequency of observing $x$ conditional on $C_j$ being the intended character; but it measures the similarity of the observation to the conceptual character $C_j$ in a way that agrees with our intuition.
Bibliography


136


