# A Bayesian Framework for Concept Learning

by

Joshua B. Tenenbaum

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1999



Author .....
Department of Brain and Cognitive Sciences
February 15, 1999

Certified by .....
Whitman A. Richards
Professor of Cognitive Science
Thesis Supervisor

Accepted by .....
Gerald E. Schneider
Chairman, Department Committee on Graduate Students

# A Bayesian Framework for Concept Learning

by

Joshua B. Tenenbaum

Submitted to the Department of Brain and Cognitive Sciences
on February 15, 1999, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

Human concept learning presents a version of the classic problem of induction, which is made particularly difficult by the combination of two requirements: the need to learn from a rich (*i.e.* nested and overlapping) vocabulary of possible concepts and the need to be able to generalize concepts reasonably from only a few positive examples. I begin this thesis by considering a simple number concept game as a concrete illustration of this ability. On this task, human learners can with reasonable confidence lock in on one out of a billion billion billion logically possible concepts, after seeing only four positive examples of the concept, and can generalize informatively after seeing just a single example. Neither of the two classic approaches to inductive inference – hypothesis testing in a constrained space of possible *rules* and computing *similarity* to the observed examples – can provide a complete picture of how people generalize concepts in even this simple setting.

This thesis proposes a new computational framework for understanding how people learn concepts from examples, based on the principles of Bayesian inference. By imposing the constraints of a probabilistic model of the learning situation, the Bayesian learner can draw out much more information about a concept's extension from a given set of observed examples than either rule-based or similarity-based approaches do, and can use this information in a rational way to infer the probability that any new object is also an instance of the concept. There are three components of the Bayesian framework: a *prior* probability distribuion over a hypothesis space of possible concepts; a *likelihood* function, which scores each hypothesis according to its probability of generating the observed examples; and the principle of *hypothesis averaging*, under which the learner computes the probability of generalizing a concept to new objects by averaging the predictions of all hypotheses weighted by their *posterior* probability (proportional to the product of their priors and likelihoods). The likelihood, under the assumption of randomly sampled positive examples, embodies the *size principle* for scoring hypotheses: smaller consistent hypotheses are more likely than larger hypotheses, and they become exponentially more likely as the number of observed examples increases. The principle of hypothesis averaging allows the Bayesian framework to accomodate both rule-like and similarity-like gen-

eralization behavior, depending on how peaked the posterior probability is. Together, the size principle plus hypothesis averaging predict a convergence from similarity-like generalization (due to a broad posterior distribution) after very few examples are observed to rule-like generalization (due to a sharply peaked posterior distribution) after sufficiently many examples have been observed.

The main contributions of this thesis are as follows. First and foremost, I show how it is possible for people to learn and generalize concepts from just one or a few positive examples (Chapter 2). Building on that understanding, I then present a series of case studies of simple concept learning situations where the Bayesian framework yields both qualitative and quantitative insights into the real behavior of human learners (Chapters 3-5). These cases each focus on a different learning domain. Chapter 3 looks at generalization in continuous feature spaces, a typical representation of objects in psychology and machine learning with the virtues of being analytically tractable and empirically accessible, but the downside of being highly abstract and artificial. Chapter 4 moves to the more natural domain of learning words for categories of objects and shows the relevance of the same phenomena and explanatory principles introduced in the more abstract setting of Chapters 1-3 for real-world learning tasks like this one.

In each of these domains, both similarity-like and rule-like generalization emerge as special cases of the Bayesian framework in the limits of very few or very many examples, respectively. However, the transition from similarity to rules occurs much faster in the word learning domain than in the continuous feature space domain. I propose a Bayesian explanation of this difference in learning curves that places crucial importance on the density or sparsity of overlapping hypotheses in the learner's hypothesis space. To test this proposal, a third case study (Chapter 5) returns to the domain of number concepts, in which human learners possess a more complex body of prior knowledge that leads to a hypothesis space with both sparse and densely overlapping components. Here, the Bayesian theory predicts – and human learners produce – either rule-based or similarity-based generalization from a few examples, depending on the precise examples observed. I also discusses how several classic reasoning heuristics may be used to approximate the much more elaborate computations of Bayesian inference that this domain requires.

In each of these case studies, I confront some of the classic questions of concept learning and induction: Is the acquisition of concepts driven mainly by pre-existing knowledge or the statistical force of our observations? Is generalization based primarily on abstract rules or similarity to exemplars? I argue that in almost all instances, the only reasonable answer to such questions is, "Both." More importantly, I show how the Bayesian framework allows us to answer much more penetrating versions of these questions: How does prior knowledge *interact* with the observed examples to guide generalization? *Why* does generalization appear rule-based in some cases and similarity-based in others? Finally, Chapter 6 summarizes the major contributions in more detailed form and discusses how this work fits into the larger picture of contemporary research on human learning, thinking, and reasoning.

To my parents, Marty and Bonnie Tenenbaum, without whom there wouldn't even be a hypothesis space,

and

to Mira, who believed in the Hazaka Principle.

# Acknowledgments

I am fortunate to have many people to thank.

My advisor Whitman Richards, for the freedom to go where my heart and mind led me, and for the guidance that kept me from squandering that freedom.

Roger Shepard, for his inspiration all along the way and his careful and critical reading of my work.

Steve Pinker, for helpful suggestions on methodology and literature, for always pushing me to clarify the goal of my work, and for providing a model of how to make formal theories of learning relevant to cognitive science.

Aaron Bobick, for his understanding of why this work is worth doing and for not pulling any punches.

Mike Jordan, for showing me – along with a generation of students – the power of uncertain knowledge.

Ted Adelson and Bart Anderson, certainly unofficial advisors, who gave helpful advice on research and career and taught me how to talk about vision.

Jacob Feldman, my "older brother" in the study of concept learning.

Bill Freeman, my first collaborator, for many lessons in both the style and content of scientific research.

Mike Tarr, for pointing me in the right direction.

The Howard Hughes Medical Institute, for its support in the form of a predoctoral fellowship.

Jan Ellertsen and Denise Heintze, for the real support.

Pat Claffey, for tolerating (and sometimes even contributing to!) the unofficial branch of the E10 library which I set up in my office.

All the participants in my experiments, and particularly the last-minute ones, for whom $10 only begins to express the depths of my gratitude.

Hany Farid, Jenny Ganger, Sharmin Ghaznavi, Stephen Gilbert, Josh McDermott, Marina Meila, John Mikhail, Chris Moore, Thanos Siapas, David Somers, Cristina Sorrentino, James Thomas, Emanuel Todorov, and Yair Weiss: true friends and col-

leagues during my time in graduate school.

Raphael Lehrer, for his friendship and his patience when I wanted to pretend I was still a physicist.

Ari Hershowitz, for his friendship and for the single most encouraging thing anyone has ever said to me, delivered at a time when I needed it the most.

Russ Burnett, for his friendship and his career choice, which helped convince me that I wasn't the only crazy one.

Fei Xu, for her friendship, encouragement, and profound influence on almost every aspect of this work.

Harriet Katz, my grandmother and most devoted reader.

Finally, those to whom this thesis is dedicated. Mira Bernstein and my parents, Bonnie and Marty Tenenbaum, have followed this work more closely than anyone over the last three years. Any scientist would be lucky to have one person in the world who loves him and also genuinely understands, appreciates and contributes to his work; I have these three. No words of thanks could possibly measure up to their support.


PSYCHE
THERE WAS A THESIS HERE
SOMEWHERE

# Contents

# Chapter 1

# Introduction: The computational problem of concept learning

The ability to learn concepts from examples is a central aspect of human cognition. Yet while every child can naturally acquire many concepts from only very limited evidence of what the concept refers to, the computational basis of this ability is still poorly understood. Consider the state of the art in machine learning, a field which has seen tremendous progress in the last 15 years. Computers can learn to tell the difference between images of the handwritten digits 2, 3, and so on, after training on thousands of labeled examples of each class (Hinton, Dayan, Frey & Neal, 1995; Simard, LeCun & Denker, 1993). Other systems can learn to detect faces in images at varying scales and positions, after training with thousands of examples of images labeled either "face" or "no face" (Osuna, Freund, & Girosi, 1997; Rowley, Baluja & Kanade, 1998). Now consider the "state of the art" in human learning: children only six years old routinely learn words that refer to coherent, but complex and overlapping, units of their worlds, such as "'dog", "fish", "pet", "Rover", and "animal". They learn these words at the remarkable rate of five or more a day, being given only a few relevant examples of how each word is used and no systematic evidence of how words are not to be used (Carey, 1978; Markman, 1989; Regier, 1996; Bloom & Markson, 1998). Computers, for all their impressive successes on certain learning tasks, don't even come close to the everyday achievements of any child.

This thesis presents a computational theory of concept learning with three intimately linked goals: first, to account for how concept learning is even possible from the limited evidence typically available to human learners; second, to explain some of the most central empirical phenomena of human concept learning; and third, to suggest how artificial systems may begin to approach the generalization capabilities of human learners. Learning theories are usually classified as *normative*, if they are meant to provide a standard for ideal or optimal learning behavior, or *descriptive*, if they are meant to provide an account of actual human learning behavior. Normative learning theories are, these days, in the realm of computer science (formerly philosophy); descriptive theories the province of psychologists (and increasingly, neuroscientists). This thesis attempts to bridge the normative and the descriptive traditions by constructing, within the accepted normative framework of Bayesian inference, a realistic theory of human learning that explains important aspects of human behavior not successfully explained in the past. This work thus contributes to the growing body of "rational" cognitive science exemplefied by Shepard (1987), Anderson (1990), and the papers in Oaksford & Chater (1998). It is *not* meant to be a mechanistic theory, to describe the details of neural systems or the processes of memory and attention that underlie concept learning. Rather, in the traditions of Chomsky (1986) and Marr (1982), it is a theory of *knowledge* – what sort of knowledge is brought to bear in learning concepts, what sort of knowledge is acquired from experience, and how that knowledge is used to generalize concepts to new situations – and it is a theory of *computation* – what is the computational problem that the concept learner must solve, what is the nature of the input, and what are the constraints necessary to guarantee a useful solution. These are the issues at the intersection of psychology and computer science (not to mention neuroscience and philosophy) that must be addressed by anyone who wants to understand the nature of concept learning in humans or machines.

I hope that my attempts to balance normative and descriptive goals will make this work of interest to readers coming from both traditions. To sum up the message for psychologists in one sentence: we should think more like computer scientists, focusing

on an analysis of the hard computational problems underlying concept learning and asking whether or not our psychological models are capable of solving these problems. To sum up for computer scientists: we should think more like psychologists, focusing on the learning situations that play the most important roles in human cognition and asking whether or not our computational models are appropriate for these situations. I think that both of these fields have the tendency to focus on problems that are more tractable than relevant, which is probably wise given how far we stand from a scientific understanding of mind. Although I will not claim to have solved the really hard problems of human concept learning, I do hope that I can make them seem at least a little bit more tractable, and thus, worthy of our attention at this stage of the inquiry.

The plan of this introductory chapter is as follows. After laying out the necessary vocabularly and background material (Section 2), I illustrate by way of a simple and concrete example – the "number concept game" – what makes human concept learning so remarkable from a computational point of view (Section 3). I argue that concept acquisition presents a version of the classic problem of induction, which is made exceptionally difficult by the combination of two requirements: the need to learn from a rich (*i.e.* nested and overlapping) vocabulary of possible concepts and the need to be able to generalize concepts reasonably from only a few positive examples. In Sections 4 and 5, I review two classic approaches to inductive inference – hypothesis testing in a constrained space of possible *rules* and computing *similarity* to the observed examples – and argue that neither one alone can provide a complete picture of even such a simple concept learning task as the number concept game. Section 6 considers how we might construct a theory that bridges traditional rule-based and similarity-based approaches to concept learning, distinguishing a *unified* theory – the target of this thesis – from *modular* theories, typified by several recently published models.

Let me also briefly outline the main contributions of this thesis (and the plan for subsequent chapters). First and foremost, I develop an understanding of how it is possible for people to learn concepts from just a few positive examples based on

the principles of Bayesian inference (Chapter 2). Then, building on that understanding, I present a series of case studies of simple concept learning situations where the Bayesian framework yields both qualitative and quantitative insights into the real behavior of human learners (Chapters 3-5). These cases each focus on a different learning domain. Chapter 3 looks at generalization in continuous feature spaces, a typical representation of objects in psychology and machine learning with the virtues of being analytically tractable and empirically accessible, but the downside of being highly abstract and artificial. Chapter 4 moves to the more natural domain of learning words for categories of objects and shows the relevance of the phenomena and explanatory principles introduced in more abstract settings for real-world learning tasks like this one. Chapter 5 considers a domain of number concepts, in which human learners possess a more complex body of prior knowledge that leads to more subtle phenomena not observed in the earlier case studies, and discusses how various classic reasoning heuristics may be used to approximate the much more elaborate computations of Bayesian inference.

In each of these case studies, I confront some of the classic questions of concept learning and induction: Is the acquisition of concepts driven mainly by pre-existing knowledge or the statistical force of our observations? Is generalization based primarily on abstract rules or similarity to exemplars? I argue that in almost all instances, the only reasonable answer to such questions is, "Both." More importantly, I show how the Bayesian framework allows us to ask much more penetrating versions of these questions, such as: How does prior knowledge *interact* with the observed examples to guide generalization? *Why* does generalization appear rule-based in some cases and similarity-based in others? Finally, Chapter 6 summarizes the major contributions in more detailed form and discusses how this work fits into the larger picture of contemporary research on human learning, thinking, and reasoning.

## 1.1 Vocabulary and Background

Practically nothing about "concepts" and "concept learning" is universally accepted by all who study them, least of all the essential definitions of these terms. In this thesis, I will begin by treating concepts as pointers from the mind to subsets of the world (Millikan, 1998). I will focus primarily on object concepts, that is, pointers to subsets of objects in the world. The subset of entities in the world to which a concept points or applies is called its *extension*. The extensions of concepts are often called *categories*. Our most basic object concepts are captured by the words (in particular, the count nouns) of our native language – "dog" points to the set of dogs, "bee" points to the set of bees – but more complex pointers – "raindrops on roses", "whiskers on kittens", "warm woolen mittens" – may serve as concepts as well.

The acquisition of even relatively simple object concepts like "dog" is undoubtedly a complex business, drawing on multiple cognitive processes over an extended period of time. This thesis focuses on the stage of *ostensive* concept learning: learning a concept's extension from examples, *i.e.* object-label pairs. Examples belonging to the concept's extension are called *positive*; those outside the extension are called *negative*. Examples may also be *implicit*, that is, dependent on inferences from other existing pieces of knowledge. For instance, a positive example of "dog" may implicitly also be a positive example of "animal" and a negative example of "cat", if we know that dogs are also animals and that no animal is both a dog and a cat.

Crucially, examples may be generated by many different processes. For the case of a child learning the concept "dog", labeled examples could be provided deliberately by a knowledgeable adult – "Here's a dog", "Here's another dog", "This one's also a dog' – or selected by the child herself in the form of queries to the adult – "Is this a dog?", "What about this one?" – or encountered at random as the child moves through her world and notices the labels adults use spontaneously. Frequently there is an asymmetry between how positive and negative examples are generated. While adults routinely label positive examples of object concepts – "See the doggie?" – they much more rarely provide spontaneous negative examples – "See that? That's not a

doggie." However, negative examples may often occur in the form of feedback on a child's mistaken identifications – "No dear, that's not a dog, that's a cat." A formal model of the process generating the learner's observations will be called a *generative model*. Differences between generative models, such as between deliberately provided samples and feedback on the learner's misidentifications, will turn out in the next chapter to be crucial for the possibilities of concept learning. This thesis focuses primarily on the case when a teacher provides the learner with a small number of positive examples sampled at random from the concept, because this is the basic situation of word learning. However, we also want to leave room for negative evidence provided in the form of corrective feedback. Hence our models need to be able to learn meaningfully from only positive evidence, but also need to be able to accomodate negative evidence when available.

## 1.2 The problem

To begin to see where the difficulties of ostensive concept learning lie, consider the following simple learning game. I have written some short computer programs, each of which checks to see if numbers satisfy a single easy-to-state arithmetical concept. The concepts are nothing fancy or tricky; some possibilities might be "X is even", "X is between 30 and 45", "X is a power of 3", "X is less than 10", and so on. Each program takes as input a natural number (*i.e.* an integer greater than zero) and returns as output either "yes" or "no", depending on whether the number satisfies some particular concept. To keep things simple, we'll assume that only numbers less than 100 are under consideration. Your job as learner is to guess what one of these programs will do from seeing only examples of its inputs and outputs. Specifically, the computer will show you a few examples of numbers which that program says "yes" to, and you will identify the other numbers you think the program will say "yes" to, and also how confident you are. Keep in mind that there is nothing special about the examples you will see – they are chosen randomly from all the numbers (less than 100) that the program would say "yes" to. Ready?

. . . The computer has chosen a program . . .

. . . and a random example of a number it says "yes" to: 16.

Now, what other numbers do you think this program will say "yes" to? 17? 5? 6? 2? 64? It's very hard to say with only one example, but it does seem that some numbers are more likely to be accepted than others. For example, to me, it seems that 17 is more likely to be accepted than 87, based on its relative proximity to 16. However, numerical proximity isn't the whole story; 32 seems to me somewhat more likely to be accepted than 30, because it shares more arithmetical properties with 16. Similarly, 4 seems to me more likely to to be accepted than 5 or 6. But these preferences are pretty vague – nothing I would want to bet on. And you may feel rather differently; maybe you think 6 is more likely to be accepted than 4, because it shares the same last digit as 16. After all, we do have very little information to go on.

Suppose the computer now generates three more random examples of numbers its program says "yes" to: 8, 2, and 64. These three examples, in addition to "16", help a lot. Now it is much clearer to me which other numbers the program will accept: 32 and 4 become much more likely to be accepted than 30 or 6, and both 17 and 87 seem quite unlikely to be accepted. Of course no firm proof has been established for any of these generalizations. Yet most readers would agree that given the four random "yes" examples of 16, 8, 2, and 64, this program seems most likely to pick out the powers of two, *i.e.* 2, 4, 8, 16, 32, 64 and so on. (If you find that you're not in agreement, let me assure you that these examples were randomly chosen from all those numbers the program accepts. Does that help?)

The major phenomena I seek to explain in this thesis are judgments of *generalization* like these. Figure 3 shows the average responses that people make to precisely these questions on an experimental survey. [1] Participants in this study were asked to rate the probability that various numbers would be accepted by the computer program, given first one example 16 (Figure 3, top row) and then three more examples

[1] This study is discussed in detail in Chapter 5.

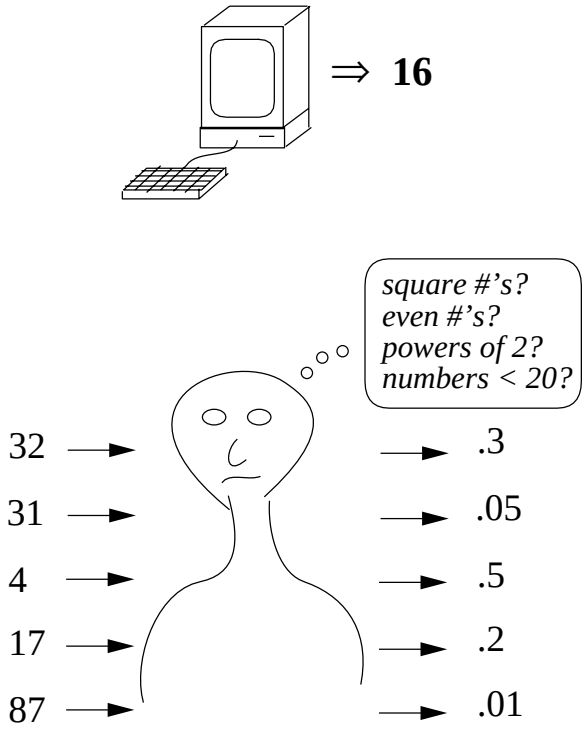



**Figure 1**

8, 2, and 64 (Figure 3, bottom row). Ratings were made on a scale of 1-7, here normalized to a probability between 0 and 1. It is clear that people's judgments are quite uncertain – but not arbitrary – given just one example, but that they quickly converge to the judgment that all and only the powers of two will be accepted after seeing just three more examples.

Now, appearances perhaps to the contrary, something quite remarkable happens inside your brain when you produce these generalization judgments. Even restricting the game to natural numbers between 1 and 100, there are more than a billion billion billion subsets of numbers that such a program could possibly have picked out and which are consistent with the observed "yes" examples of 16, 8, 2, and 64.[2] That's

[2]To be precise, there are $2^{100}$ subsets of numbers between 1 and 100, and each positive example cuts in half the number of logically consistent subsets (*i.e.* those containing all the positive examples).

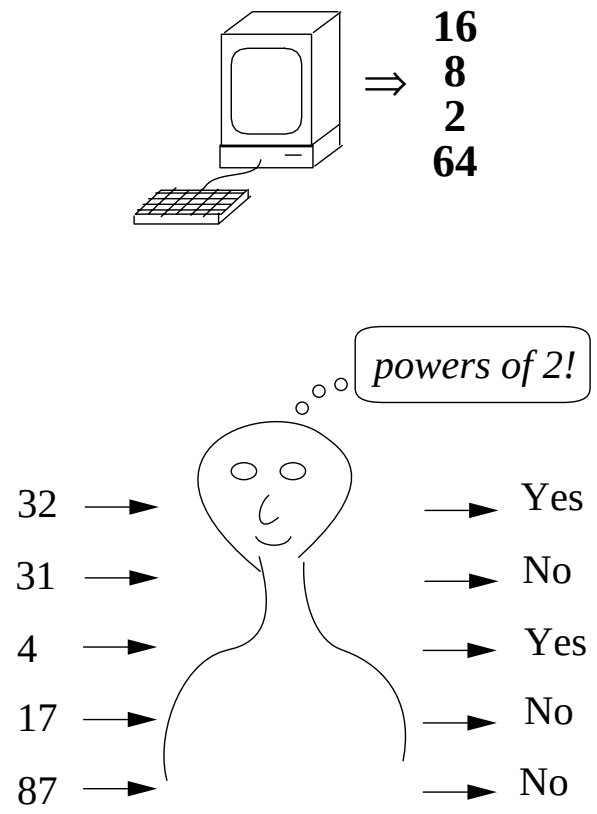



**Figure 2**

a billion for every second since the beginning of the universe. These subsets include not just *all powers of two*, but *all even numbers*, or *all numbers less than 100*, not to mention possibilities like *all powers of two and also 37*, or *all powers of two except 32*. Yet despite this practically infinite range of possibilities, you feel to some degree confident that you can identify numbers in the one subset that this program actually does pick out. Moreover, this confidence comes after seeing just four random examples known to be in that set – *positive examples* of the computer's concept – and *no* negative examples, *i.e.* numbers known to be not in that set. Trying to explain

This is because for every subset $s$ containing a particular number $x$, there is exactly one other subset that contains all of the same numbers as $s$ except $x$. Thus, after four examples, we are left with $2^{100-4} = 2^{96}$ logically consistent subsets.

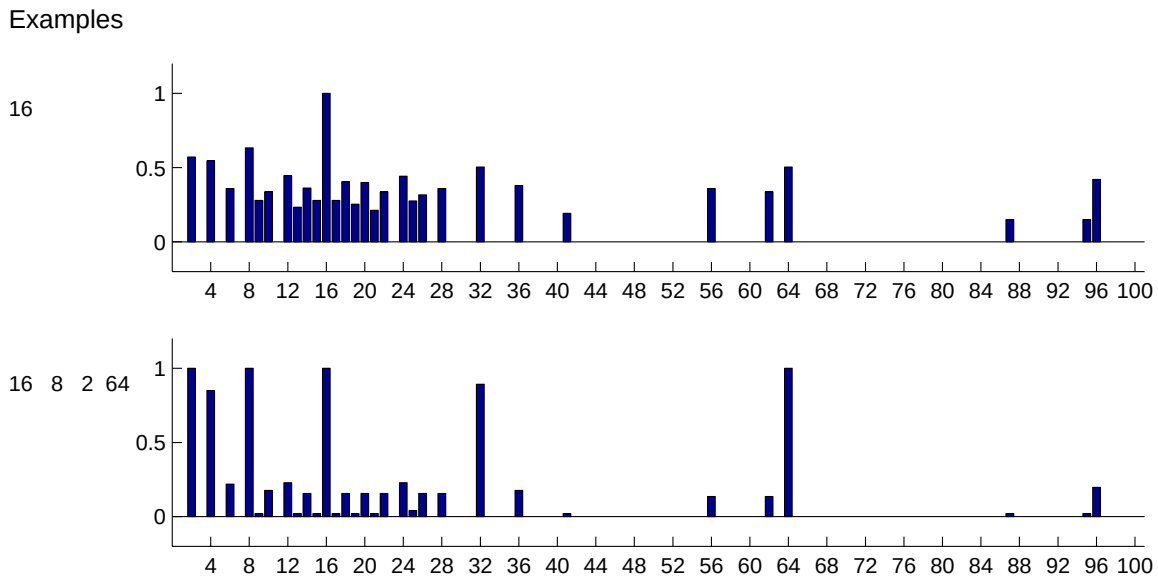



**Figure 3**

away this inference as merely "common sense" only highlights how mysterious the processes involved really are. This is an instance of the classic problem of induction, which has been at the center of philosophical attention since Bacon's *New Organon* (1620) and at the center of philosophical controversy at least since Hume's famous *Treatise* (1739). It will not be dismissed lightly.

Perhaps this game of guessing numerical concepts can be brushed off on psychological grounds, as too artificial or impoverished a task; perhaps it should be treated as *only* a puzzle for philosophers. However, the number-program game deserves our attention because it highlights the essential inductive challenge underlying many more natural – but much more complex – concept learning tasks. To see this, consider first the case of learning words for object concepts, perhaps the most basic form of human concept learning. Specifically, let's consider what sort of inferences are required of a child learning to use a word like "dog".

There is some subset of the entities in the world – the set of dogs – which the word "dog" refers to. Again, this subset is called the *extension* of the concept labeled by "dog". In coming to use the word "dog" competently, the child has to learn to identify all and only those entities which fall into its extension. Now, if an expert were

to correctly label every object in the world as either "dog" or "not a dog", and the child could somehow observe and remember all of these labelings, then the problem would be solved. Or if there were a simple definition of "dog" that picked out all and only the dogs, and if somebody could communicate this definition to the child, then too the problem would be solved. But neither of these learning modes is realistic for most words.

In general, the child must infer the correct extension of "dog" from only a few experiences of dogs in conjunction with the linguistic label "dog" (Figure 4). After having heard the word "dog" applied to, say, a particular beagle, a particular labrador, and a particular dalmation, the child ideally should be willing to apply that term not to only those three individual animals, nor to only beagles, golden retrievers and dalmations, nor to all animals nor all living things, but to all and only the dogs. Receiving negative feedback on her mistaken identifications, as when she calls a cat "dog" and then hears "no, cat", will certainly be helpful in establishing precisely the extensions of new concepts. However, extensive negative feedback is definitely not necessary – and could hardly be helpful! – before she is willing to call only a restricted, nonarbitary set of things in the world "dog" (Carey, 1978; Markman, 1989; Regier, 1997; Bloom & Markson 1998).

Now consider a computer system that must learn concepts from interacting with human users. It will face essentially the same inferential ambiguities. A good example is a computer system for interactive scene analysis (Barrow, Bolles, Garvey, Kremers, Lantz, Tenenbaum & Wolf, 1977) that learns to identify images or image regions satisfying a particular visual concept such as *trees*, *sand*, or even *dogs*. [3] I might like such a system to automatically label all regions of *trees* or *dogs* in an image or set of images, given a few regions that I have labeled as examples of *trees* or *dogs*. Ideally, the computer should be just as productive yet selective as the child above in generalizing these concepts to new image regions, because that is how I, the human user, am used to interacting with other human learners. Providing more than a few

[3]For contemporary examples of such a system, see Minka & Picard (1997) and De Bonet & Viola (1997).

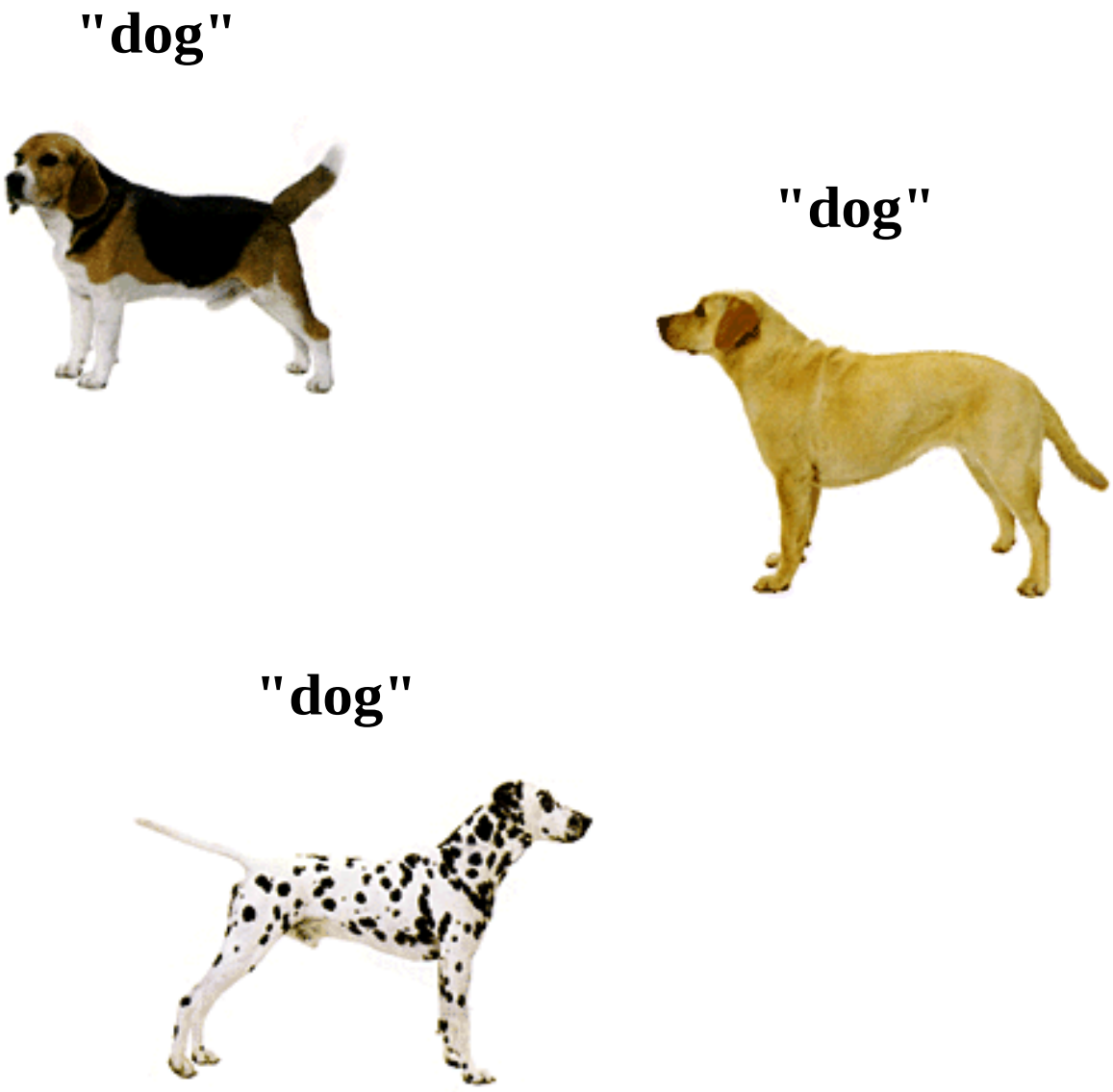



**Figure 4**

examples of *trees* regions quickly gets tedious, and providing informative negative examples, *e.g.* good examples of *non-trees*, may be difficult for the average user.

My claim is that the enterprises of learning to use words, learning scene concepts, or guessing numerical concepts have at their core the same computational problem: they all require the learner to infer *how far* and *in what ways* to generalize from only a *limited number* of *positive examples*. Inferring "how far" to generalize a concept means that the learner must choose the appropriate generalization from many possibilities that are nested inside each other: *e.g.* { *labradors*, *dogs*, *animals* }, { *pine trees*, *trees*,

*plants* }, or { *powers of two*, *even numbers*, *all numbers* }. Inferring "in what ways" to generalize means that the learner must choose from possible generalizations that are partially overlapping: *e.g.* { *dogs*, *pets*, *four-legged animals* }, { *trees*, *leaves*, *green plants* }, or { *powers of two*, *square numbers*, *cube numbers* }. Somehow, from all of these reasonable ways to generalize, the learner must determine the one subset that contains all and only the instances of the relevant concept. The restriction to a "limited number of positive examples" means that the learner must be able to generalize informatively – if not perfectly – after seeing any number of examples drawn only from the concept's extension. Again, this is not to say that negative evidence is never available or helpful, but that it is generally of secondary importance, *i.e.* provided as negative feedback on the learner's own mistaken identifications of instances of a concept, rather than as the primary evidence on which these initial tentative identifications are based.

Perhaps *the* major result of computational studies of learning in the last few decades is this: the broader and more complex the range of possible concepts a learner is capable of acquiring, the greater the number (and/or the stronger the kind) of examples required to learn successfully (Geman, Bienenstock & Doursat, 1992; Vapnik, 1995; Valiant, 1984; Kearns & Vazirani, 1994). Human concept learning poses a particularly thorny version of this challenge, combining the need to learn from a rich (*i.e.* nested and overlapping) vocabulary of possible concepts with the need to learn each concept more or less from only a few positive examples. Dealing with either of these requirements on its own would be much less difficult. Seeing most of the possible positive examples, or seeing good negative examples, would cut down the nested and overlapping possibilities the learner must usually contend with. Restricting the possible concepts to only *disjoint* (non-nested, non-overlapping) sets of objects would allow any concept to be learned trivially from just a single positive example, because each object would belong to at most one possible concept! How such a range of potential concepts can be acquired from such limited evidence is the great mystery of human concept learning, the great challenge of machine concept learning.

## 1.3 The rule-based approach to induction (and why it doesn't work as a theory of concept learning)

The classic problem of induction that underlies concept learning has a classic solution, which in various forms is one of the major positions on learning in cognitive psychology, machine learning, and philosophy of science. For reasons that will become clear, I will call this cluster of approaches to inductive inferences *rule-based* approaches. In this section, I describe the basic rule-based approaches and illustrate why they are not adequate to explain how people learn concepts from just a few positive examples.

### 1.3.1 Simple hypothesis elimination with *a priori* constraints on the hypothesis space

There are two main ideas behind the classic rule-based approach to inductive inference. The first idea is that learning proceeds via *hypothesis elimination*, otherwise known as the hypothetico-deductive method. That is, the learner considers various hypotheses about what the extension of the concept could be, and eliminates those which are not consistent with the examples observed. This idea alone is not worth much as a theory of learning, because an infinite number of general hypotheses will always be consistent with any finite set of observations; how is the learner to generalize? Enter the second idea of the rule-based approach: the process of hypothesis elimination is guided by a *hypothesis space* subject to *a priori constraints*. That is, instead of considering all logically possible hypotheses consistent with a set of examples, the learner considers only a much smaller subset of hypotheses – those in his hypothesis space – which according to his prior knowledge are natural candidates for being the concept's extension. Generalization from limited evidence becomes practical because most of the possible ways to generalize are never even considered by the learner. The hope is that, with strong enough constraints on the hypothesis space and a reasonable number of observations, the learner can rule out all but one hypothesis as inconsistent

with the evidence, and the remaining hypothesis will determine his generalizations. This way of looking at concept learning was first proposed by Hovland (1952), in the form of a "communication analysis" of the teacher-learner exchange, and developed in the last twenty years primarily in the field of machine learning (Mitchell, 1979; 1997).

Let's see how the notion of a hypothesis space of natural rules would be useful in explaining concept learning in the number concept game. In particular, we want to explain the intuition that, given the random "yes" examples of 16, 8, 2, and 64, the program probably accepts all and only the powers of two, as opposed to any of the other logically possible subsets of numbers it could accept.

Why do we infer that the program probably accepts all and only the powers of two, *i.e.* $\{2, 4, 8, 16, 32, 64, \ldots\}$, and not the powers of two plus one other arbitrary number, *e.g.* $\{2, 4, 8, 16, 32, 37, 64, \ldots\}$, or all of the powers of two except one, *e.g.* $\{2, 4, 8, 16, 64, \ldots\}$? The classic rule-based approach answers this question by saying that we have (implicitly or explicitly) adopted a certain hypothesis space of candidate extensions for the concept, containing not all logically possible extensions but only those that seem natural in the context of learning a simple arithmetical concept. This hypothesis space includes what seem like mathematically natural hypotheses, such as *all powers of two*, but does not include seemingly unnatural hypotheses like *all powers of two, and also 37*, or *all powers of two, except 32*. Then, if we decide how to generalize by searching our hypothesis space for a candidate concept that is consistent with the observed examples, these psychologically bizarre but logically possible generalizations simply never come up. In other words, *before we have seen any examples of the concept at all*, our possible generalizations are constrained by our prior beliefs about the natural concepts in this domain.

The importance of a strong hypothesis space in guiding generalization from limited evidence was first recognized by nineteenth century British philosophers trying to understand the logical basis of inductive inference in science. Like many ideas in the philosophy of science, this one was stated first and best by William Whewell:

> But supposing the Facts to be adequately observed, they can never be

> combined into any new Truth, except by means of some new Conceptions, clear and appropriate, such as I have endeavored to characterize. When the observer's mind is prepared with such instruments, a very few facts, or it may be a single one, may bring the process of discovery into action.
>
> (Whewell, *Novum Organon Renovatum*, 1858)

In this century, the idea that meaningful inductive inference requires some prior knowledge about which hypotheses are more natural than others has become well entrenched in philosophy through the work of Goodman (1955) and those who picked up the challenge of his "new riddle of induction". Philosophers following Goodman distinguish between "projectible" hypotheses, which receive confirming support from the observation of evidence consistent with them, and "nonprojectible" hypotheses, which do not receive support from consistent observations. The classic case is the contrast between green and grue, where grue means "green if observed before Jan. 1, 2000, and blue if observed thereafter." The hypothesis "All emeralds are green" receives support from the observation of green emeralds, but the hypothesis "All emeralds are grue" – just as consistent with the observation of all green emeralds seen up to the present day – does not seem intuitively to be confirmed by the same evidence. Goodman attributed the difference to a difference in projectibility – green is projectible, grue is not – and philosophers ever since have been trying to figure out what this really means.

From its mid-1800's Cambridge origins and its renaissance in Cambridge, Mass. in the mid-1900's, the notion of a hypothesis space constrained by prior knowledge has spilled over into modern cognitive psychology, linguistics, and machine learning, and now occupies a central place in those fields' major paradigms. The earliest empirical studies of human concept learning (Bruner, Goodnow & Austin, 1956; Shepard, Hovland & Jenkins, 1961; Hunt, 1962; Bower & Trabasso, 1964) fall primarily into this rule-based approach. Cognitive development researchers have seized on the importance of constraints in guiding the child's process of word learning and concept acquisition (Keil, 1979; Osherson, 1978; Markman, 1989; Bloom, in press; Shipley,

1993). For instance, it has been proposed that children's first concepts are arranged hierarchically in a taxonomic tree, with no concept a subset of more than one other concept (Keil, 1979), and that their first words map onto nodes in this taxonomic tree (Markman, 1989). Linguists have repeatedly invoked hypothesis space constraints to explain how children can acquire the grammar of their native language from only positive evidence (Chomsky, 1986; Pinker, 1995; Gibson & Wexler, 1994; Osherson, Stob & Weinstein, 1986). Machine learning researchers have likewise stressed the learner's need for "inductive bias," prior knowledge that somehow restricts the space of possible concepts that can be learned in order to make hypothesis elimination a computationally tractable strategy (Mitchell, 1982; Haussler, 1988). The most popular sources of inductive bias in machine learning take the form of limitations on the complexity of concept descriptions (*e.g.* the number of terms in a logical formula) – a version of Ockham's razor (Valiant, 1984; Kearns & Vazirani, 1994). In fact, the machine learning literature has produced a number of proofs that meaningful generalization would be impossible without such biases (e.g. Mitchell, 1980, or Watanabe, 1985).[4]

With such a wide base of precedents, constraints on conceptual naturalness are the obvious place to start looking for a solution to the computational problem of concept learning. Moreover, it is clear that prior knowledge plays an essential role in how people generalize concepts. For example, making any inferences at all about the powers of two clearly requires the prior knowledge (or at least the prior disposition to think) that this is a class worth paying attention to. A child who only knows the rudiments of arithmetic may think the set $\{2, 4, 8, 16, 32, 64, \ldots\}$ just as strange as the set $\{2, 4, 8, 16, 64, \ldots\}$; to him, both are just "random" collections of even numbers. At the other extreme of prior knowledge, mathematicians know the triangular numbers $\{1, 3, 6, 10, 15, \ldots n(n-1)/2 \ldots\}$ have special significance in many arithmetical problems. Given the positive examples 10, 3, 6, and 21, a mathematician on the lookout for such a sequence might be quick to generalize to 15 and 28, but not 16

[4]That is, the probability that *any* entity $x$ belongs to *any* concept $C$ would be $1/2$, unless $x$ had been specifically labeled as a positive or negative instance of $C$.

and 27; a normal adult most likely would not have these preferences.

So, no one can doubt that hypothesis space constraints play an important role. The question is, are they the whole story? And the answer to that is just as clearly "no". Looking back at the number concept example shows us what is missing. The observed examples 16, 8, 2, and 64 are consistent with *more than one* a priori natural hypothesis, yet we infer that a single hypothesis is significantly more likely than the others to be the true concept. Why do we infer that the program picks out all and only the powers of two, as opposed to all even numbers, or all numbers less than 100? It seems that a priori – before we have seen any examples of numbers that the program accepts – all three of these classes are reasonably natural candidates for what such a simple program might pick out. Certainly, none of these hypotheses is nearly as bizarre as the class of all powers of two except 32. [5] Moreover, the even numbers seem to be, if anything, more familiar and natural a class, a priori, as the powers of two. Thus prior beliefs about the relative naturalness of these different classes cannot explain why we settle with some confidence on just *one* of these hypotheses, all and only the powers of two, after observing examples which are logically consistent with all of them.

The real problem with this rule-based approach to induction, as I have portrayed it so far, is that it tries to force inductive inference into the rigorous but rigid mold of deductive inference. Theoretically, this strategy can work if we have a limited enough hypothesis space and a large enough supply of examples to guarantee that only one hypothesis will survive elimination by inconsistent data. Then, given the premises that the true concept belongs to the hypothesis space and that the true concept is consistent with the observed examples, it follows deductively that the one remaining hypothesis is the true concept. However, this strategy is bound to fail

[5] One good test for a "bizarre" or "unnatural" hypothesis is whether the hypothesis fails to be compelling even after all of its instances have been observed as examples. After seeing the examples 16, 8, 2, 64, 4, 32, it is pretty compelling that the program accepts all and only powers of two. Likewise after seeing all the even numbers as examples, or all the numbers less than 100. However, after seeing the examples 16, 8, 2, 64, and 4, the hypothesis *all powers of two except 32* still fails to be compelling; it starts to become compelling after seeing perhaps 15 examples, 16, 8, 2, 64, 2, 8, 4, 64, 16, 4, 2, 64, 8, 8, 16. Likewise for *all powers of two and also 37*: it's not compelling after seeing just 16, 8, 2, 64, 37, 4, 32, but starts to become so after $16, 8, 2, 64, 37, 4, 32, 8, 64, 16, 37, 32, 32, 4, 2, 64, 8, 37, 16, \ldots$.

for most cases of human concept learning, which are essentially inductive. Recall the point made above, that the natural candidate extensions for concepts are often nested (*e.g.* { *animal*, *dog*, *labrador*, *Rover* }) or partially overlapping (*e.g.* { *woman*, *parent*, *mother* } ). Thus any very small set of positive examples of a concept is likely to belong to more than one candidate extension, and some further criterion – beyond a priori naturalness – will be required to distinguish among these possible generalizations. In many situations, such as the healthy levels task considered in Chapter 3, the learner may even have to face an *infinite* set of nested and overlapping generalizations, each consistent with the given set of examples, and each an a priori natural candidate for the concept to be learned. These situations make the starkest case for some additional source of information in concept learning.

### 1.3.2 Ranked hypothesis elimination

Perhaps this simple rule-based view is too simple; perhaps it can be saved if the initial constraints are allowed to be soft, not hard. That is, rather than merely allowing or disallowing candidate hypotheses, the constraints induce a ranking of all possible hypotheses in order of a priori conceptual naturalness, and the learner chooses the highest ranked hypothesis that is consistent with the examples. Such a view of constraints has been advanced in psycholinguistic accounts of language acquisition (Pinker, 1984; Wexler & Manzini, 1987; Berwick, 1985). It is also probably closer to how most cognitive developmentalists envision the role of constraints than the strict in-or-out picture presented above. I will call the idea of an a priori ranking of hypotheses the *ranked rule-based* approach to induction, to be distinguished from the *simple rule-based* approach given above.

Under the ranked rule-based view, to justify the preference for *powers of two* over *even numbers* after 16, 8, 2, and 64, we would have to assume that *powers of two* is placed significantly higher than *even numbers* in the a priori ranking. But then we are at a loss to explain why generalization from one example, 16, or two examples, 16 and 8, is so different from generalization from the four examples 16, 8, 2, and 64. If the preference for *powers of two* over *even numbers* is explained solely by an

a priori ranking, then it should be just as strong after only one or two examples – equally consistent with both hypotheses – have been observed. However, this does not agree with intuition: *powers of two* seems to have little (if any) advantage over *even numbers* after just 16 has been observed, but seems to become rapidly more compelling after we have seen just a few more consistent examples.

### 1.3.3 Flexible hypothesis preferences

Perhaps we can amend the ranked rule approach by allowing the rankings of hypotheses to vary from their a priori settings, depending on the particular examples observed. In this case, many hypotheses including *powers of two* and *even numbers* would have very similar (perhaps equal) rankings a priori. After we have seen one example, these rankings might change slightly but not enough to give any one hypothesis a clear advantage over the others. After we have seen a few more examples, however, the rankings change enough to clearly favor one hypothesis – *powers of two* – and we now generalize strictly according to that rule. The story sounds good, but it is no more than a redescription of the phenomena unless we can provide a mechanism for how (and an explanation for why) the rankings of hypotheses change with the observed examples. This is in a sense where this thesis is headed. Work in coming chapters will be devoted to understanding exactly how the weights of hypotheses depend on the observed examples, and how generalization behavior depends on those weights.

The number concept task illustrates another, deeper problem with rule-based approaches to concept learning; namely, the very character of generalization seems to shift as we go from one example to four, in a way that the rule-based approach cannot explain. After we have seen the four examples 16, 8, 2, and 64, one hypothesis seems clearly better than any others and generalization is based strictly on whether or not numbers conform to the rule *powers of two*. However, when we have only seen the one example 16, and many hypotheses seem more or less reasonable, generalization to other numbers seems more graded, based on their overall similarity to 16. 4 seems more likely to be accepted than 5, 17 seems more likely to be accepted than 87, and so on. The idea that concepts are always generalized in accordance with

the best fitting rule – or any one rule at all – just seems wrong in cases like this one. Numerous experiments in the psychological literature have documented the importance of similarity-based generalization in cognition (see Goldstone, 1994, for a review).

In sum, neither the simple rule-based account of induction via a constrained hypothesis space nor the ranked rule variant can account for two important aspects of learning on the number concept task. First, candidate extensions which are roughly equal in a priori naturalness, and which have similar degrees of plausibility given only one example, may take on very different plausibilities after a few more examples – consistent with both! – are observed. Second, how we generalize is not always determined by the single best hypothesis. When many hypotheses receive roughly equal plausibility, *e.g.* when just a single example has been observed, a gradient of generalization based on overall similarity seems more natural than sharp, all-or-none generalization based on any single rule. These phenomena will arise whenever the hypothesis space is rich enough, and the number of examples few enough, to ensure that the observed data cannot eliminate all but one candidate hypothesis. That is to say, they will occur *virtually always* when people are learning natural concepts from just a few positive examples.

What we need in a model of human concept learning is a model that can manage the uncertainty of multiple consistent hypotheses, a model that explains why some consistent hypotheses seem increasingly more likely than others as more examples are observed and how generalization can be based on multiple hypotheses of (perhaps) varying plausibilities. The Bayesian framework for concept learning proposed in this thesis addresses these two core needs. Most generally, the essence of Bayesian inference is the use of probability theory to represent and reason about uncertainty. In the case of concept learning, the relevant uncertainty is the learner's uncertainty about the concept's extension – *i.e.* the presence of multiple consistent hypotheses – given only a few positive examples. The Bayesian framework for concept learning incorporates the classic notion of an a priori constrained hypothesis space and adds to it the sophisticated inferential machinery necessary to decide how to generalize when

the observed examples do not uniquely determine a single consistent hypothesis, thus forming the basis for a model of how people learn and generalize concepts from very limited positive evidence.

## 1.4 The similarity-based approach (and why it doesn't work either)

While the classic approaches to inductive inference have been rule-based, there is an alternative tradition in induction that has been extremely influential in psychology and machine learning, and that seems to succeed precisely where rule-based approaches fail as accounts of human concept learning. Before turning to the Bayesian approach to concept learning that is the heart of this thesis, it is worth considering what these alternatives have to offer. These approaches reject wholesale the idea that generalization is guided by an a priori hypothesis space of candidate rules. Rather, they assume that the learner's prior knowledge and/or innate endowment equips him with a primitive sense of *similarity*, and that the learner generalizes a concept from observed examples to a new object on the basis of that new object's degree of similarity to the examples.

### 1.4.1 Concept learning based on pairwise similarity

To be precise, we will take similarity to be a function SIM($y \rightarrow x$) from pairs of objects $x$ and $y$ to some space of degrees of resemblance; these degrees of similarity will be assumed to be real-valued (*e.g.* SIM(Bill $\rightarrow$Hillary) = .8, SIM(Bill $\rightarrow$Monica) = .4, SIM(Bill $\rightarrow$Ken) = .05, where 1 denotes complete similarity – "everything in common" – and 0 denotes complete dissimilarity – "nothing in common"), but could also be ordinal-valued (*e.g.* SIM(Bill $\rightarrow$Hillary) > SIM(Bill $\rightarrow$Monica) > SIM(Bill $\rightarrow$Ken)). The similarity function is potentially asymmetric, that is, SIM($y \rightarrow x$) – the similarity of $y$ to $x$ – is not necessarily equal to SIM($x \rightarrow y$).

Historically, the similarity-based view of induction has its basis in Hume's doctrine

of association by resemblance. According to Hume (1739), resemblance is one of the three main forces that establishes the "connexion of ideas", along with contiguity (in time) and cause-and-effect. William James (1890/1981) elevated association by similarity, over association by contiguity and other means of inductive inference, as "our chief help towards noticing those special characters of phenomena, which, when once possessed and named, are used as reasons, class names, essences, or middle terms...." (pp. 971-972). In the twentieth century, the most notable advocate for similarity as the foundation of induction has been W. V. Quine. In numerous writings over the last forty years, Quine (1960, 1969, 1995) has argued that an innate similarity space is all the guide a human infant has on his earliest and most fundamental inductive adventures. This animal sense of similarity is the product of natural selection, and as such can be expected to be a generally adaptive bias in acquiring concepts in the natural world. Wittgenstein (1953), too, is associated with the similarity-based view, through his arguments that many everyday concepts are not held together by a common essence shared by all category members, but rather by a network of "family resemblances". In legal philosophy, generalization on the basis of similarity to previous cases is a standard and accepted method of reasoning from established precedents to novel legal situations (Levi, 1949).

In psychology and machine learning as well, similarity has often been taken as the foundation of learning and reasoning. Behaviorists trumpeted similarity as the major force behind "stimulus generalization", whereby a conditioned response could be elicited not only by stimuli physically identical to the conditioned stimulus, but also to a lesser degree by stimuli similar on a relevant physical dimension (e.g. tones of similar pitches, lights of similar brightnesses). Most contemporary formal models of classification, too, incorporate some kind of similarity space as an explanatory construct (Shepard, 1964; Medin & Schaffer, 1978; Smith & Medin, 1981; Nosofsky, 1986, 1992; Ashby, 1992; Estes, 1994). In particular, the most successful exemplar theories of classification (Medin & Schaffer, 1978; Nosofsky, 1986) assume that the probability with which a subject will generalize a category label from a set of examples $X = \{x_1, \ldots, x_n\}$ to a novel object $y$ is some simple function (such as the average, sum

or maximum) of the similarities $\mathrm{SIM}(y \rightarrow x_i)$ of $y$ to each of the $x_i$. In machine learning, similar exemplar-based models have also become some of the most popular tools (Mitchell, 1997; Aha, 1997). Perhaps the most robust and consistently successful of all pattern recognition techiques is also the simplest: *nearest neighbor (NN)* classification (Duda & Hart, 1973). NN classification is based entirely on a primitive similarity function, which is used to assign each new object to the class that contains the most similar previously encountered object.

Despite the clearly established utility of similarity in models of classification, its value in explaining how humans learn and generalize concepts from just a few positive examples is not so clear. Virtually all similarity-based models in both psychology and machine learning were developed for discriminative learning tasks, in which positive and negative examples play equal and essential roles in guiding generalization. As a consequence, these models *require* negative examples of a concept in order to generalize in any meaningful way – quite unlike human concept learners. Appendix A goes through this argument in detail. On these grounds, many of the most important similarity-based models of classification learning from cognitive psychology – Kruschke's (1992) ALCOVE model, Anderson's (1991) "rational model", and various adaptive network models (*e.g.* Gluck & Bower, 1988; Gluck & Myers, 93; Estes, 1994) – fail to account for how people can infer, given only a few positive examples, which new entities a concept is likely to pick out.

Nonetheless, the construct of similarity does seem to explain certain aspects of concept learning and generalization that rule-based approaches do not. Recall that a deep problem with rule-based approaches was their inability to capture generalization behavior that did not conform to an all-or-none pattern. In the number concept game, a single rule seems quite salient when we have seen all four examples 16, 8, 2, and 64; but when we have seen just one example of the program's target set, *e.g.* 16, no single rule provides a very compelling case for generalization. Rather, the intuition is that we generalize to new numbers based on how similar they are to 16. A similarity-based theory of concept learning can capture this intuition directly, as follows: given a positive example $x$ of a concept $C$, the learner judges the probability that a new

object $y$ belongs to $C$ to be proportional to its similarity to the observed example, $\mathrm{SIM}(y \rightarrow x)$.

### 1.4.2 Can similarity to a *set* of examples account for rule-like generalization?

What about cases where people's generalization behavior seems more naturally described by a rule, as after we have seen the four examples 16, 8, 2, and 64? In order to discuss similarity-based generalization from more than one example of a concept in a rigorous way, we need to extend the definition of similarity from a pairwise relation $\mathrm{SIM}(y \rightarrow x)$ to a setwise relation $\mathrm{SIM}(y \rightarrow \{x_1, \ldots, x_n\})$ – the similarity of a new object $y$ to the *set* of $n$ observed examples $\{x_1, \ldots, x_n\}$). There is no generally accepted model of how people evaluate the similarity of one object to a set of objects, although there are a number of proposals in the psychological literature. All the major candidates suggest that $\mathrm{SIM}(y \rightarrow \{x_1, \ldots, x_n\})$ is a relatively simple function of the pairwise similarities $\mathrm{SIM}(y \rightarrow (x_i)$. Three popular proposals are:[6]

- **total similarity:** $\mathrm{SIM}(y \rightarrow \{x_1, \ldots, x_n\}) = \sum_i \mathrm{SIM}(y \rightarrow x_i)$.
- **average similarity:** $\mathrm{SIM}(y \rightarrow \{x_1, \ldots, x_n\}) = \frac{1}{n} \sum_i \mathrm{SIM}(y \rightarrow x_i)$.
- **maximum similarity:** $\mathrm{SIM}(y \rightarrow \{x_1, \ldots, x_n\}) = \max_i \mathrm{SIM}(y \rightarrow x_i)$.

At first, it seems that these similarity-based models could describe apparently rule-based cases of generalization as well. For instance, after seeing the examples 16, 8, 2, and 64, the number 4 becomes more likely to be accepted than it seemed to be after just the example 16. Because 4 is probably more similar to 8 and to 2 than to 16, it is plausible that all three setwise measures of exemplar similarity – total, average, and maximum – increase with the three additional examples. Thus the increase in generalization makes sense under any of the three exemplar models.

[6]Some references. Total similarity: Nosofsky (1986). Average similarity: Ashby & Leola-Reese (1995). Maximum similarity: Goldstone (1994), Osherson et al. (1990).

However, there are other cases where only one of three similarity models seems to match our intuitions. The number 32 also becomes much more likely to be accepted after 8, 2, and 64 are observed than after only 16 was observed. However, because 32 is not substantially more similar to any one of 8, 2, or 64 than it is to 16, only its total exemplar similarity increases, not its average or maximum exemplar similarity. Hence only the total similarity model seems appropriate in this case.

The number 14 seems much *less* likely to be accepted after 8, 2, and 64 are observed than after only 16 was observed. Because 14 is probably at least as similar to 16 as it is to 8, 2, or 64, only its average exemplar similarity can decrease with these three additional examples, not its maximum or total similarity. Thus only the average similarity model seems appropriate in this case.

Perhaps trivially, the number 16 seems *equally* likely – *i.e.* 100% likely – to be accepted after 16, 8, 2, and 64 are observed and after only 16 is observed. Clearly 16 is more similar to itself than to 8, 2, or 64, hence only its maximum exemplar similarity is equal before and after these three additional examples; its total exemplar similarity *increases* and its average exemplar similarity *decreases*. So in this case, only the maximum similarity model is appropriate.

In short, none of these similarity models can explain all cases of apparently rule-based generalization. Worse than that, there are some cases which cannot be explained by any of these models. Consider the probability that 5 will be accepted given the example 16, versus the probability that 5 will be accepted given the examples 16, 8, 2 and 4. Intuitively, generalization to 5 decreases with these additional examples, but pretty clearly 5 is at least as similar – and probably more similar – to 8, 2, and 4 as it is to 16. Thus all three similarity models should predict an *increase* in generalization, counter to intuition.

### 1.4.3 Making similarity more flexible

Even in the face of phenomena like these, many theorists remain committed to the idea of similarity as the basis of concept learning and generalization (Goldstone, 1994; Nosofsky, 1986; Medin & Florian, 1995). The solution, they would claim, is that

the similarity function SIM($y \to x$) should not be treated as a psychological primitive fixed by a priori knowledge. Rather, our sense of similarity is *flexible* across different contexts, weighting different aspects of stimuli differently depending on the examples that have been observed. Suppose we assume for concreteness that SIM($y \to x$) is just a weighted count of the features common to $x$ and $y$ (Tversky, 1977; Shepard & Arabie, 1979). In the context of number concepts, these features might be properties of numbers such as *divisible by two*, *less than 10*, and so on. Each feature receives a weight reflecting how much it contributes to the sum total of similarity between two numbers that share it. Under this model, similarity is flexible to the extent that the weights of these features can change.

This flexible similarity approach would explain the difference in generalization from 16 alone versus 16, 8, 2, and 64 as follows. Given just the one example 16, we have no particular reason to weight the property of being a power of two much more than the property of being an even number, or any other property, so generalization is determined by the sum of many factors that comprise the "overall" similarity of a number to 16. But after we have seen 16, 8, 2, and 64, the property of being a power of two – because it is shared by all of these stimuli – becomes much more salient than other properties. Now the computation of SIM($y \to 16$) (along with the other quantities SIM($y \to 8$), SIM($y \to 2$), and SIM($y \to 64$), on which SIM($y \to \{16, 8, 2, 64\}$) depends) places far greater weight on the feature *power of two* than on any other property of $y$; consequently, other powers of two will be highly similar to these examples, and non-powers of two will be highly dissimilar. Thus generalization will appear to follow a simple rule even while it's still really a similarity computation.

The power of flexible similarity to model human concept learning and generlization is obvious here. By allowing the weights of features to change, apparently rule-based generalization can be accomodated as a "special case" of similarity-based generalization, and we only need one mechanism for concept learning. The danger of this approach should also be obvious: it's completely circular! Why does the feature *power of two* receive greater weight than the feature *even number* – and much greater weight after 8, 2, and 64 are observed – even though both features are shared by all

four examples? Replace "feature" with "hypothesis" and "shared by" with "consistent with", and we're back to the same dilemna that plagued the rule-based approach to concept learning discussed in the previous section!

Frustration with this potential for circularity led Goodman (1972; see also Watanabe, 1985) to declare flexible similarity a sham, a scandal that strips the construct of similarity of any real explanatory power. More recently, Goldstone and colleagues (Goldstone, 1994; Medin, Goldstone, & Gentner, 1993) have tried to reclaim a flexible but constrained sense of similarity as a valuable explanatory tool. Several computational models of classification learning incorporate flexible similarity metrics in the form of variable feature weights (Kruscke, 1992; Aha & Goldstone, 1992; Shanks & Gluck, 1994). However, these are discriminative learning models (see the appendix to this chapter), not concept learning models. To learn, they require both positive and negative examples of a category, as well as feedback on their classification mistakes. They adjust their feature weights to maximally discriminate the set of positive training examples from the set of negative training examples. But this doesn't say anything about how people can learn which features are important from seeing just a few positive examples of a concept.

To sum up, a similarity-based account of human concept learning requires two things which current models don't have on offer. First, we need a principled account of flexibility, how the weights of features in a similarity computation change depending on the examples observed. Second, we need a principled way to compute the similarity of a new object to a *set* of examples. Above, we saw three proposals for setwise similarity functions based on pairwse similarities to examples (total, average, or maximum). Is the right way to generalize by similarity given by one of these formulae, or some other procedure?

From the point of view of the similarity theorist, these two goals are the two goals of this thesis. Recall that at the end of the last chapter, I described the goals of this thesis from the perspective of rule-based approaches to inductive learning: to provide an account of how the rankings of possible hypotheses for a concept's extension change from their a priori settings as examples are observed, and to explain

how people generalize when *more than one* reasonable hypothesis is consistent with the examples. If it's not already clear, these two statements of goals are just two ways of saying the same thing, first in terms borrowed from the rule-based tradition and then in terms borrowed from the similarity-based tradition.

## 1.5 Two strategies for building a more complete theory

Let's recap the findings of the previous two sections. We've seen how the two major traditions in inductive inference suggest two different approaches to modeling human concept learning, based on rules or similarity. Each approach specifies a basic algorithm for generalizing a concept from examples: applying an abstract rule or computing the similarity of one object to another. Each approach also specifies some a priori knowledge that constrains the generalization algorithm: an a priori ranking of hypotheses or a priori weights on the features of stimuli. Looking at our concrete example of the number concept game, we saw different kinds of generalization behavior, some that seemed best described as similarity-based (generalizing from the single example 16) and some that seemed best described as rule-like (generalizing from the four examples 16, 8, 2, and 64). Each basic approach works best under certain ideal conditions, – when there is a single clearly best rule to apply, or when generalizing by similarity to a single observed example – and these conditions are usually mutually exclusive. That is, when we have seen only one example of a concept, there will generally be more than one reasonable rule that could pick out the concept's extension; when one rule seems clearly better than any other, we will generally have seen more than example. Hence no simple model of generalization based on the ideal of applying a single abstract rule or computing similarity to a single example can hope to describe the whole course of human concept learning.

### 1.5.1 Strategy 1: a unified theory

Given this diagnosis, there are essentially two strategies for how to build a more complete theory of concept learning that would combine the virtues of traditional rule-based and similarity-based approaches. The first strategy is to aim for a *unified* theory that explains both "rule-like" and "similarity-like" generalization behavior with a single mechanism. Such a theory could itself look nothing like either rule- or similarity-based models as traditionally conceived – for example, it could look like a neural network, – or it could be based on an extension of one of these approaches. In the last two sections, I argued that either traditional approach could potentially handle the full range of concept learning phenomena if extended in two ways: first, to provide a more sophisticated generalization algorithm based on integrating *multiple* consistent rules or computing similarity to a *set* of examples; second, to provide for flexible constraints on generalization – flexible hypothesis rankings or feature weights – that allow apparently rule-guided or similarity-guided behavior to emerge from a single model depending on the observed examples.

This is the strategy I pursue in this thesis. I develop a probabilistic extension of the traditional rule-based approach, based on the principles of Bayesian inference. Specifically, the basic elements of the Bayesian learner's hypothesis space are candidate rules for picking out a concept's extension, but the learner maintains a probability distribution over those rules instead of choosing only a single rule to represent each concept. By integrating the predictions of multiple rules consistent with a set of examples, the learner can generalize concepts based on a gradient of similarity to those examples (as in Figure 3, top row). However, when the learner's probability distribution is concentrated on a single hypothetical extension, exactly the same procedure of integrating over the hypothesis space leads to apparently all-or-none, rule-governed generalization (as in Figure 3, bottom row). The Bayesian framework thus contains both rule-like and similarity-like generalization as special cases, offers a unified explanation of both sorts of behavioral phenomena, and explains *why* generalization appears rule-based or similarity-based in any particular situation.

A full exposition of the Bayesian framework will be the subject of Chapter 2; readers curious to see what the theory really looks like may want to skip ahead now. For the rest of this section, let me briefly try to justify why the unifying Bayesian account takes rules rather than similarity as its starting-off point – its basic representation of concepts – and why a unified account, as opposed to a modular one, is worth pursuing at all.

There are intuitive reasons for starting with rules, as well as highly technical ones. Intuitively, abstract rules seem fundamental to our concept of "concept", in a way that similarity to exemplars does not. Throughout the Western philosophical tradition, from the ancient Greeks to the twentieth century, a major goal has been to formulate definitions or rules that pick out precisely the referents of our concepts. What is man? The featherless biped. What is knowledge? Justified true belief. Granted, this sort of conceptual analysis is often a difficult – or even impossible – task. Nonetheless it has seemed to be *the* task worth undertaking, to many of the greatest minds of the last two thousand years. Contemporary philosophers by and large recognize the futility of such defintions but still act as though they're the only game in town for conceptual analysis. Goldman (1986) states this position explicitly:

> But given the longstanding failure of philosophers (or lexicographers) to devise accurate definitions of this kind for very many (if any) words, it is doubtful whether such analyses are indeed possible.... In the meantime, however, I will follow the working practice of trying to give necessary and sufficient conditions for important epistemological terms. This can be regarded as a first approximation to a better way of treating meaning, if such a better way can be devised. (pp. 38-39).

The conceptual work of the natural sciences has followed similar paths, in quest of increasingly more precise definitions to pick out the true units of the mathematical, physical, and biological worlds. In the social sphere, our dictionaries and legal codes are monumental attempts to define – in rules – the concepts behind how we think we use words and how we think we ought to live together. In cognitive psychology, the

so-called "classical theory" that dominated early work on concept learning (Bruner, Goodnow & Austin, 1956) took it for granted that concepts were mentally represented in the form of rules. While this view no longer holds sway as it once did, it continues to be tremendously influential as the "alternative" hypothesis to much contemporary work. Even the pioneer of the prototype theory of concepts, Eleanor Rosch, accepts the extensional side of the rule-based view of concepts: "A category exists whenever two or more distinguishable objects or events are treated equivalently" (Mervis & Rosch, 1981). Rules are simply our attempts to describe these equivalence classes of objects as precisely as possible.

Last and certainly not least, the intuition that concepts in general are or ought to be rule-based is also shared by the average person who is not a philosopher, physicist, lawyer, or cognitive scientist engaged in the professional pursuit of these defintions. Naive subjects (both adults and children), asked whether it is a matter of fact or opinion whether something was a dog, responded "fact" to an overwhelming degree (Kalish, 1998; Armstrong, Gleitman & Gleitman, 1983). Medin & Ortony (1989) described this phenomenon as "psychological essentialism": even if we don't know the true rule that picks out instances of a natural kind, we believe that there is some underlying essence that all members of the kind share, which science has or could potentially discover. An essentialist position, whether folk or scientific, is a rule-based one; to believe in the existence of some dog essence is to believe that $X$ is a dog if and only if $X$ has dog essence.

My point in bringing up each of these examples is not to prove that our concepts really *are* describable in terms of rules; on the contrary, all of these attempts to formalize concepts in rule form are either highly artificial or, more or less, failures! Rather, the point is to illustrate that to a great many people over a great many years of human history, rules for applicability have seemed to be at the core of conceptual knowledge, or at least as close as we can get to it. I take this as good prima facie evidence that there is something important about rules worth preserving in a theory of concept learning. Clearly, from the fact that no one – despite exhaustive efforts – has succeeded in capturing a natural concept in rule form, we can infer that rules

are not a *sufficient* basis for our concepts. That's why I propose to extend rule-based approaches, rather than embrace them and rest. But the most intriguing thing about rules or definitions is how popular they remain despite their consistent failure to capture the true structure of anything interesting in the real world. I take this as the *best possible* evidence that our brains are set up to conceive of the world in this way, in terms of classes defined by rules. If it turned out that rule-based concepts in fact described the world successfully, their ubiquity would be unsurprising. But their consistent *failure* to describe the world, combined with their ubiquity in both scientific and lay thought, provides a "poverty of the stimulus"-like argument for their primary psychological reality.

More intuitive support for the primacy of rule-based concepts comes from looking at how rules and similarity intersect in language use. To a first approximation, words are binary symbols; they are either used or not. If we want to comment on the dogness of $X$, there is only one word we can use: "dog". We do not have multiple words – such as "dog", "dogg", "doggg", "dogggg", and so on – to express varying degrees of dogness or similarity to a dog prototype; we do not have a continuous family of speech sounds between /dog/ and /kat/ to express the relative similarity of an animal to the dog prototype versus the cat prototype. The fact that we frequently hedge or qualify our words – "OK, technically speaking, a chihuahua is a dog; but a german shepherd, now that's a real dog!" [7] – has been taken as good evidence that there is more to our concepts than just rules of application (Lakoff, 1972). It's a point well made and well taken, but crucially, it's rule-based statements that are being hedged to make way for variations in similarity, not the other way around.

The position that a unified theory of concept learning ought to begin from a rule-based approach is also supported by several technical arguments. Fodor (1998) has argued that only rule-like representations can support what he sees as the most important cognitive function of concepts: composition and combination. Rules – in the form of necessary and sufficient conditions – can be combined with each other

[7] Or Paul Hogan's famous line from the movie *Crocodile Dundee*: "That's not a knife.... Now *that's* a knife!"

using the AND, OR, and NOT operators of classical logic. Moreover, these logical combinations preserve the compositionality of meaning: $X$ is a brown cow if and only if $X$ is brown AND $X$ is a cow. Other proposals for conceptual structure – stored exemplars, feature statistics or theories – do not seem able to combine compositionally. The Bayesian framework to be developed in this thesis preserves the compositionality of classical rule-based approaches in cases where it is appropriate, and also makes clear when such simple compositionality is not appropriate.

Perhaps the best reason to proceed from rules towards similarity is that the classic explanatory models of similarity have themselves taken precisely this approach – if under somewhat different names! We have already seen that any similarity-based account of concept learning requires a flexible but constrained notion of similarity. This means that the similarity between two objects cannot be a primitive relation in the mind, but must derive from some more fundamental knowledge that allows similarity to vary across contexts while still imposing some structure on how it is allowed to vary. Tversky's (1977) well-known contrast model of similarity, mentioned above, assigns this role to features. Objects are represented as sets of feature elements, each feature is assigned a weight, and the similarity between two objects is a function of the total weight of their common features minus the weight of their distinctive features. But what are features if not rules in disguise? Instead of thinking of objects as sets of features, we can think of features as sets of objects; the two pictures are completely equivalent (mathematically speaking, they are dual). Once features are seen as picking out sets of objects, they are no different from rules. Saying that two numbers 6 and 4 share the features *less than 10* and *even* is no different from saying that they both belong to the set of numbers less than 10 or the set of even numbers; it's also no different from saying that the rules "X is less than 10" or "X is an even number" classify both 6 and 4 as positive instances. Hence any feature-based account of similarity is really just a rule-based account in disguise, where two objects are similar to the extent that they fall under the extensions of the same rules.

Another influential tradition of similarity models is based on the geometry of metric spaces, otherwise known as the *multidimensional scaling (MDS)* approach.

In MDS, we model the pairwise similarity relations over a set of objects by representing each object as a point in a low-dimensional metric space and defining the similarity of two objects to be some decreasing function of the distance between their corresponding points in this "psychological" space (Shepard, 1980). The flexibility of similarity is modeled by stretching or shrinking a set of coordinate axes for this space (Nosofsky, 1986). Now, it is a non-trivial general finding of the MDS approach that similarity is very well described by an exponentially decaying function of distance in psychological space (Shepard, 1987). The theory that Shepard offers to explain the universality of the exponential decay can – like Tversky's model – be viewed as a rule-based account of similarity. Shepard proposes that the similarity of two objects reflects the probability that they both belong to an arbitrarily chosen "consequential region", a subset of the psychological space that corresponds to the extension of a natural kind concept. Just as we did for Tversky's features, we can view each of these consequential regions as the extension of a rule. Similarity under Shepard's theory then measures the probability that two objects will fall under the extension of an arbitrarily chosen rule.

I will have much more to say about Shepard's (1987) theory later in this thesis (Chapters 2 and 3) – indeed, it was one of the principal inspirations for the Bayesian framework developed here. The important point to take away from this discussion is that the two classic explanatory pictures of similarity – as a measure of feature overlap or distance in psychological space – can both be viewed as deriving similarity from a computation over more primitive rule-based representations. This idea will figure centrally in the chapters to come, as we see how similarity-based generalization in concept learning derives from probabilistic operations over an essentially rule-based generalization mechanism. By formulating concept learning as Bayesian inference over a base of rule-like hypotheses, we are able not only to account for both rule-guided and similarity-guided generalization behavior, but also to explain the rational basis of each kind of behavior, and to predict – within a given domain of objects as well as across different domains – the conditions under which each will occur.

### 1.5.2 Strategy 2: a modular theory

The alternative to unifying rule-based and similarity-based approaches under one theory is to develop a *modular* theory, which acknowledges the independent importance of both rule and similarity-based generalization by hypothesizing two distinct concept learning modules, one for each kind of processing, as well as perhaps a third controller module for mediating between the other two. The idea that separate rule-based and similarity-based computations both participate in human concept learning is probably in fact the modal view in contemporary cognitive science. Analogous proposals have been made for a number of other aspects of cognition, ranging from reasoning (Sloman, 1996) to language acquisition (Pinker, 1991). Several authors writing in a recent special issue of *Cognition* devoted to the topic (Sloman and Rips, 1998) have argued that rules and similarity are (the!) two fundamental operating modes for human minds. In AI, similar ideologies underly the development of hybrid systems that incorporate various combinations of rule-based and "fuzzy", "case-based" or "similarity-based" reasoning components (e.g. Sun, 1995). Most relevantly for this thesis, the categorization judgments of human subjects often seem to reflect both rule-based and similarity-based knowledge (Armstrong, Gleitman & Gleitman, 1983; Osherson et al., 1990; Smith & Sloman, 1994). People can be taught to classify objects based on either rule-based or exemplar/similarity-based procedures (Allen & Brooks, 1991), with different brain systems thought to be involved in each case (Smith et al., 1998). Finally, several authors in the mathematical psychology community have recently proposed hybrid models of classification learning based on something like a combination of rule-based and similarity-based modules (Erickson & Kruschke, 1998; Ashby et al., 1998; Nosofsky, Palmeri & McKinley, 1994; Nosofsky & Palmeri, 1998).

Despite the volume of work already in support of a modular account, there are still good reasons to develop a unified framework as I do in this thesis. A little exercise in neuroscience-fiction shows us why. Suppose that we *knew* for certain that there were two distinct modules in the brain devoted to concept learning, one specialized in applying abstract rules and the other in computing similarity to exemplars. Even

then, we would still want to know the answers to many questions that only a unified theoretical treatment can address:

- *Why* is the processing behind concept learning divided into distinct brain modules? Does it reflect the essential nature of the computational problem of concept learning, the need to achieve an efficient implementation in neural tissue, or merely an accident of evolution? Would any machine concept learning system that hoped to match human performance have to be modular in this way?

- Why are there *two* modules, and not three? Or seven, or seventeen?

- Why do those two modules implement rule-based and similarity-based computations, as opposed to some other functions?

- How does each module learn from examples, and why? Are there common statistical principles that explain how learning works in both modules?

- How do the rule and similarity modules interact? When and why is each module dominant in behavior? How do they work together to achieve a common purpose? Should we expect to find a third module that coordinates their activity, or should they be able to decide amongst themselves which module will control behavior in any given situation?

In reality, we certainly don't know yet that these two modules exist in the brain. [8] However, there do seem to be two distinct behavioral modes corresponding to rule- and similarity-based generalization, and all the above questions about brain modules apply just as well to understanding dissociations in behavior. *Why* are there distinct behavioral modes? Why two? Why these two? How does each mode work? How do they interact? The Bayesian approach developed in this thesis offers a unified framework for answering all of these questions; no strictly modular theory can.

[8] Some studies have reported that rule-based and similarity-based computations activate different brain regions (Smith et al., 1998; Ullman, Corkin, Coppola, Hickok, Growdon, Koroshetz & Pinker, 1997), although the evidence is fairly preliminary.

Moreover, no modular computational model has yet addressed the fundamental question that launched this thesis: how can people learn concepts from only a few positive examples? It's all very well to say that a similarity module handles generalization from the single example 16, while a rule module handles generalization from the four examples 16, 8, 2, and 64, but we still want to know why the rule module settles on the hypothesis *powers of two*, instead of *even number* or some other candidate. Existing modular models (Erickson & Kruschke, 1998; Ashby et al., 1998; Nosofsky & Palmeri, 1998) are all models of discrimination or classification learning, requiring both positive and negative examples for meaningful generalization, and thus can say nothing about this case (see Appendix A for details). In the Bayesian framework, in contrast, a single statistical principle explains not only why *powers of two* dominates *even number* given the examples $\{16, 8, 2, 64\}$, but also why the rule "module" dominates the similarity "module" after those four examples but *not* when only 16 was observed. For the phenomena of concept learning from limited positive evidence, a unified Bayesian theory provides both a deeper and a more parsimonious explanation than any strictly modular theory has to date.

Finally, it should be clear that the modular and unified modeling strategies need not be mutually exclusive – if, that is, we believe in multiple levels of explanation a la Marr (1982). A unified theory may offer the best explanation of the computational basis of concept learning, while a modular theory may better describe how those computations are implemented cognitively or in the brain. Heit (1998) makes essentially this argument in defense of a Bayesian account of category-based induction, which is closely related to the Bayesian framework for concept learning developed here. In Chapter 5, I'll come back to this idea that rules and similarity are unified at a computational level but implemented in a modular fashion, when I discuss psychologically plausible heuristic approximations to Bayesian concept learning.

# Chapter 2

# A solution proposed

In the first chapter, I posed the challenge of concept learning in computational terms, as an instance of the classic problem of induction. Working within the number conceptdomain, I then showed the limitations of the two classic approaches to inductive inference (testing hypotheses in a space of possible rules and computing similarity to exemplars), which each account for different aspects of how people generalize beyond the observed examples of a concept. Finally, I argued that a more complete theory of concept learning should aim to unify rule- and similarity-based approaches, taking rules as the basic representation, as opposed to taking similarity as the primitive notion, or constructing a strictly modular theory with distinct rule- and similarity-based computations. This chapter follows up on that plan, developing a *probabilistic rule-based* approach to concept learning based on the principles of Bayesian inference. At the risk of wearing out its welcome, I'll continue to use the number concepttask in this chapter's exposition, for the sake of continuity with the first chapter. The scope of the Bayesian framework is far broader, however, as I will illustrate with applications to several different domains in the following three chapters. (Readers who become impatient with this artificial example are advised to glimpse every so often at the pictures of real objects in Chapter 4.)

I should note from the outset that the framework presented here incorporates and synthesizes many ideas already present individually in the literatures of cognitive psychology, machine learning, Bayesian statistics, and philosophy of science. A

comprehensive catalog of those influences is impossible; however, let me acknowledge the strongest influences of which I am conscious: in psychology, Roger Shepard, Amos Tversky, Whitman Richards, and Jacob Feldman; in machine learning, Satosi Watanabe, David Haussler, Tom Mitchell, and Geoff Hinton; in Bayesian statistics, Sir Harold Jeffreys, E. T. Jaynes, Judea Pearl, Glenn Shafer, and David McKay; in philosophy of science, Karl Popper and Paul Horwich. I should also mention Stephen Muggleton, whose recent work in Bayesian inductive logic programming (Muggleton, preprint) independently reaches conclusions about learning from only positive data that are similar to mine in some important respects. I defer a detailed discussion of these and other authors' related work until the end of this chapter, after I have laid out my own framework and its implications.

Recall from last chapter the big question of inductive concept learning, and the two gaps in the classic rule-based account which the Bayesian framework is meant to bridge. The big issue is the problem of generalization: how does the learner infer, from a small set of positive examples of a concept, which other objects are likely to fall under that concept? Or, in the number conceptexample, why do we infer that, given the random "yes" examples of 16, 8, 2, and 64, the program probably accepts all and only the powers of two? In the rule-based approach, the learner adopts a *hypothesis space* of candidate extensions for the concept, which does not contain all logically possible extensions, but only a much smaller subset that are psychologically natural in the relevant context. In the context of the number concepttask, these hypotheses might include common mathematical classes such as *all powers of two*, but not seemingly unnatural classes like *all powers of two, and also 37*, *all powers of two, except 32*, and so on. The natural hypotheses may also be ranked in order of a priori plausibility. Upon observing the examples 16, 8, 2, and 64, all of the a priori natural but now inconsistent hypotheses (*e.g. odd numbers*, *multiples of four*) are eliminated, and the learner chooses the top-ranked remaining hypothesis as his rule for generalization: *e.g. powers of two*.

The two big questions left unanswered by this story are as follows. First, why should the learner settle on one hypothesis, *e.g. all powers of two*, over others that

are equally consistent with the observations $\{16, 8, 2, 64\}$ and that seem equally or more natural a priori, such as *all even numbers* or *all numbers less than 100*? Second, how should the learner generalize when no single hypothesis is clearly more compelling than all others, as after the one example 16? These two problems are not pathologies of artifical domains like the number concepttask, but occur frequently in natural settings such as word learning – whenever the possible concepts in the learner's hypothesis space have nested or overlapping extensions and the learner must generalize from only one or a few positive examples of the concept.

## 2.1 Statistical intuitions and generative models

Before diving into the full Bayesian formalism, let me sketch an intuitive picture of the proposed solution. Consider why, given the examples 16, 8, 2, and 64, we infer that the program picks out only the powers of two, as opposed to all even numbers, or all numbers less than 100. We cannot just explain this away on the basis of prior beliefs about conceptual naturalness, *i.e.* by assigning the hypothesis *all even numbers* a much lower a priori ranking than *all powers of two*, as we did for *all powers of two and also 37*. This would dictate, counter to intuition, that the hypothesis *all powers of two* should be just as strongly preferred over *all even numbers* after only one example, *e.g.* 16, had been observed, or even before any examples were observed! Rather, a preference for *powers of two* that emerges only *after* several examples have been observed must be intrinsically statistical, something like a drive to detect and avoid *unexplained coincidences* in the relation between concepts and their examples. It's true that the examples $\{16, 8, 2, 64\}$ are compatible with a program that accepts all even numbers. In that case, however, it would be a very suspicious coincidence indeed that no even numbers which were *not* powers of two appeared in the first four examples, if the examples really were drawn randomly from all acceptable numbers. The importance of avoiding "suspicious coincidences" in concept learning was first stressed by Feldman (1997), and has roots in the study of genericity phenomena in vision (Witkin & Tenenbaum, 1983; Koenderink, 1979; Binford, 1981; Lowe,

1985; Freeman, 1994), Barlow's theories of cerebral cortex (Barlow, 1985, 1996), and Garner's information theoretic analysis of category structure (Garner, 1962, 1974).

Crucially, the intuition that $\{16, 8, 2, 64\}$ would be a very "suspicious" set of even numbers rests on the learner having a particular *generative model* of his observations, *i.e.* an assumption about the process generating the examples he sees. A simple generative model appropriate in this case is that the examples are *independent random samples* from the true concept. That is, each example is chosen at random from all of the numbers (less than or equal to 100) that the program accepts. If the program picks out all and only the even numbers, then each example could with equal probability be any one of the 50 even numbers less than or equal to 100. If the program picks out all and only the powers of two, then each example could with equal probability be any one of the six such numbers less than 100. [1] On the other hand, suppose that instead of randomly sampled examples, the learner himself chooses the examples and the program merely labels them as "yes" or "no", in accordance with its concept. Or suppose that the examples were chosen by spinning a 100-sided spinner, or rolling two 10-sided dice. Now, if the learner (or the spinner or the dice) happened to come up with 16, 8, 2, and 64, and the program said "yes" to all four, what inferences could we make about the computer's concept? The hypothesis that the program accepts only the powers of two, rather than all even numbers, is not nearly as strong now as when the computer generated the examples, *i.e.* when the examples were taken to be sampled at random *from the concept*. The fact that all the examples happen to be powers of two is still surprising, but the surprise is focused on the learner (or the spinner or the dice) that happened to choose four powers of two in a row, and not on the program, which merely labeled all the examples as "yes".

These intuitions will be formalized below. For now, I just want to point to the

[1] This is also called *sampling with replacement*. We can imagine that the acceptable numbers are written on slips of paper in a box, and each example is generated by drawing (*sampling*) one slip at random from the box. After the example is recorded, that slip is *replaced* in the box so that all numbers have the same chance of being drawn for the next example. In some concept learning situations, perhaps including this one, it might be more appropriate to consider sampling *without replacement*, *i.e.* where the slips are not replaced after each example. For instance, if we expect that when the teacher shows us $n$ examples, they will always be $n$ *distinct* examples, then sampling without replacement is the right model.

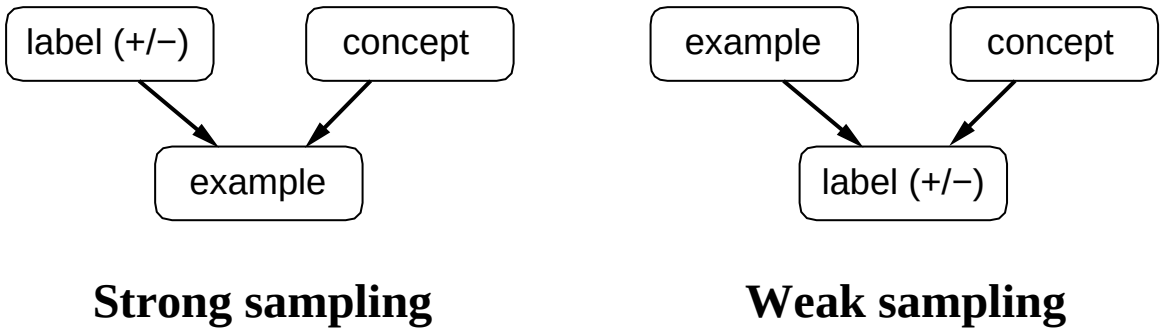



**Figure 1**

importance of generative models for the possibilities of concept learning, and to identify these two alternative models. I will call the model of examples randomly sampled from the true concept *strong sampling*, and models in which the examples are sampled independently from the concept and merely labeled by it *weak sampling* models. Of course other generative models are possible, but these two are some of the simplest and most natural, and will be the main focus in this thesis. Figure 1 depicts the difference between the strong and weak generative models in graphical format. The nodes represent the (random) variables relevant to the learning situation; the arrows, intuitively, represent direct causal influences (formally, conditional independence relationships; Pearl, 1988).

## 2.2 The independent need for prior beliefs and generative models

The traditional source of inductive leverage in rule-based approaches is a strong base of prior beliefs about the possible extensions of concepts, embodied in the a priori ranking of hypotheses. This hypothesis space makes generalization from finite experience possible, but at a severe cost: any concept not already in the hypothesis space is not learnable at all, no matter how much data is observed! Now that we are introducing an additional source of inductive leverage, a generative model which allows us to rank hypotheses on the basis of the actual *observed data* rather than on unverifiable a priori grounds, it might be tempting to think that this *statistical* constraint is all

we need, and that we can dispense altogether with the a priori hypothesis space and the strict limitations on learnability it imposes.

In fact, prior knowledge and generative models provide complementary sources of constraint, both of which are almost always necessary to learn concepts from small sets of positive examples. Just as prior beliefs do not help much to explain our preference for the hypothesis *powers of two* over the hypothesis *even numbers*, the drive to avoid unexplained coincidences under a particular generative model does not help to explain our preference for *powers of two* over strange hypotheses like *powers of two except 32*. In fact, the coincidence argument slightly favors the strange hypothesis in this cases! If the program really accepts *all* powers of two, then it is a minor coincidence that one such number, 32, did not occur among four random examples all less than 100. Of course, this could well happen by chance, but appeals to accident should be the last resort of any explanation. The hypothesis that the program accepts all powers of two *except* 32, on the other hand, explains this non-occurence as no coincidence at all. Yet, we still do not think it likely that this strange hypothesis truly describes the program's behavior; its slightly better explanation of a marginal coincidence is more than outweighed by its high *prima facie* implausibility. [2] Thus, our inference that the program accepts all and only the powers of two can only be justified by the interaction of two forces, one pulling us towards our prior belief in simple concepts and the other pushing us away from unexplained coincidences.

## 2.3 The Bayesian framework for concept learning

I hope by now it is clear that the interaction between prior beliefs (about the extensions of possible concepts) and a preference to avoid unexplained coincidences (in the relation between concepts and their examples) lies at the heart of the human competence in concept learning. In the remainder of this chapter, I will show that Bayesian inference provides a natural framework for formalizing and understanding

[2] The balance could shift, however, as the magnitude of the coincidence increases. What if we saw these 15 random examples of numbers (less than 100) that the program accepts: 16, 8, 2, 64, 2, 8, 4, 64, 16, 4, 2, 64, 8, 8, 16?

this interaction. In particular, the Bayesian framework addresses these three crucial problems of knowledge in concept learning: [3]

1. *Content:* what constitutes the learner's (uncertain) knowledge about which entities a new concept refers to?

2. *Acquisition:* how can the learner acquire that (uncertain) knowledge from the evidence provided – one or more positive examples of the concept, and possibly (but not necessarily) negative examples?

3. *Generalization:* how does the learner use that (uncertain) knowledge to generalize the concept, that is, to decide whether a particular new entity falls under the concept?

I stress the *uncertain* character of the learner's knowledge as a reminder that concept learning is a species of *inductive* inference. In contrast to deductive inferences, which yield conclusions that follow *necessarily* from their premises, inductive inferences yield conclusions that are merely *probable* given their premises. [4] Managing uncertainty must thus be an essential aspect of any inductive system. The formalism of Bayesian inference provides a principled and coherent framework for representing and manipulating the uncertainties that arise in concept learning, based on the probability calculus. The rest of this section gives a high-level overview of a Bayesian framework for concept learning in terms of the three questions posed above and illustrates it on the number-program game; a richer and more technical picture of the theory will be the subject of Chapters 3-5 (and Appendices B and C).

[3] These questions are patterned after Chomsky's organizing question for the study of language. See, *e.g.* Chomsky (1986).

[4] For a good discussion on the differences between inductive and deductive inference, see Skyrms (1986).

### 2.3.1 What constitutes the learner's knowledge about a new concept's extension?

We assume, as in the classic rule-based approach, that our knowledge about which entities a concept refers to can be represented in terms of a *hypothesis space* $\mathcal{H}$ of possible extensions of the concept. [5] The elements of this space $\mathcal{H}$ are our basic hypotheses about what the concept could refer to; each hypothesis corresponds to some candidate extension for the concept, some subset of the relevant universe of entities. In the case of learning the word "dog", these hypotheses might include various subsets of animals that could plausibly be the set of all dogs. In the case of the number-program game, the hypothesis space could include all the previously-mentioned possible subsets of numbers that the program might accept – *powers of two*, *even numbers*, *powers of two except 32*, *numbers less than 100*, etc. – and many others as well.

After we have seen a sequence $X = \{x_1, \ldots x_n\}$ of $n$ examples of a concept $C$, our knowledge about $C$'s extension in the classic rule-based approach consists of a subset of $\mathcal{H}$ corresponding to those hypotheses which are consistent with the examples in $X$. This subset of hypotheses in $\mathcal{H}$ consistent with $X$ has been called the *version space* of $X$ (Mitchell, 1979), which we will write as $\mathcal{H}_X$. As we see more examples and rule out more hypotheses in $\mathcal{H}$ as inconsistent with our observations, the version space $\mathcal{H}_X$ shrinks and we reach increasingly more certain states of knowledge about $C$'s extension.

In the Bayesian framework, our knowledge about $C$'s extension is more fine-grained and thus more powerful than just $\mathcal{H}_X$. Instead of monitoring only whether a hypothesis is consistent or not, we maintain a *probability distribution* over the hypothesis space, indicating how likely each hypothesis is to be the true extension of $C$. We denote this probability as $p(h|X)$, the probability that hypothesis $h$ is the true extension of the concept, given the $n$ examples $\{x_1, \ldots x_n\}$ that we have seen so far.

[5] I certainly do not mean to imply that we can represent *all* of our knowledge about a concept in this way, merely the information that is, to first order (*i.e.* ignoring inter-concept interactions), relevant for identifying new instances.

These probabilities are numbers between 0 and 1 reflecting our degree of belief in $h$; $p(h|X)$ is near 1 only if we are quite confident that $h$ is the true extension, near 0 if we are quite confident that $h$ is not the true extension, and somewhere in between if we are somewhat uncertain. As probabilities, these degrees of belief are normalized to sum to 1 over the hypothesis space $\mathcal{H}$:

$$\sum_{h \in \mathcal{H}} p(h|X) = 1.$$

This means that if we believe one hypothesis is almost certainly the concept's true extension, then we must be almost certain that all the other hypotheses are not. In order to assign probabilities over the hypotheses, they must be assumed to be *mutually exclusive* and *exhaustive* descriptions of the true state of affairs. In other words, one and only hypothetical extension is assumed to be the true extension of the concept.[6]

Note that if we assign zero probability to any hypothesis that is not consistent with one or more examples, then the distribution $p(h|X)$ contains all the information in the version space $\mathcal{H}_X$. But in fact, it contains much more information, in marking some consistent hypothesis as more likely than others.

Figure 2 illustrates a simple hypothesis space for the number concept game and a probability assignment over that space, given that we have seen the one positive example 16, *i.e.* $X = \{16\}$. For present purposes, the exact probability values and how they are determined are not important; I take up these issues in the next section. Also, for the sake of concreteness, we have considered only a small set of simple mathematical hypotheses: odd numbers, even numbers, square numbers, multiples of $j$ for $3 \leq j \leq 10$, numbers ending in the digit $j$ for $1 \leq j \leq 9$, and powers of $j$ for $2 \leq j \leq 10$. These are meant to be representative only, and many more hypotheses than those shown would be considered by a real human learner or a realistic computer simulation. (The simulation of number conceptlearning in Chapter 5 uses more than

[6] If it seems odd to call two hypotheses like *powers of two* and *even numbers* mutually exclusive (because every power of two is an even number), remember that *powers of two* is really short for "the extension of $C$ consists of all and only the powers of two."

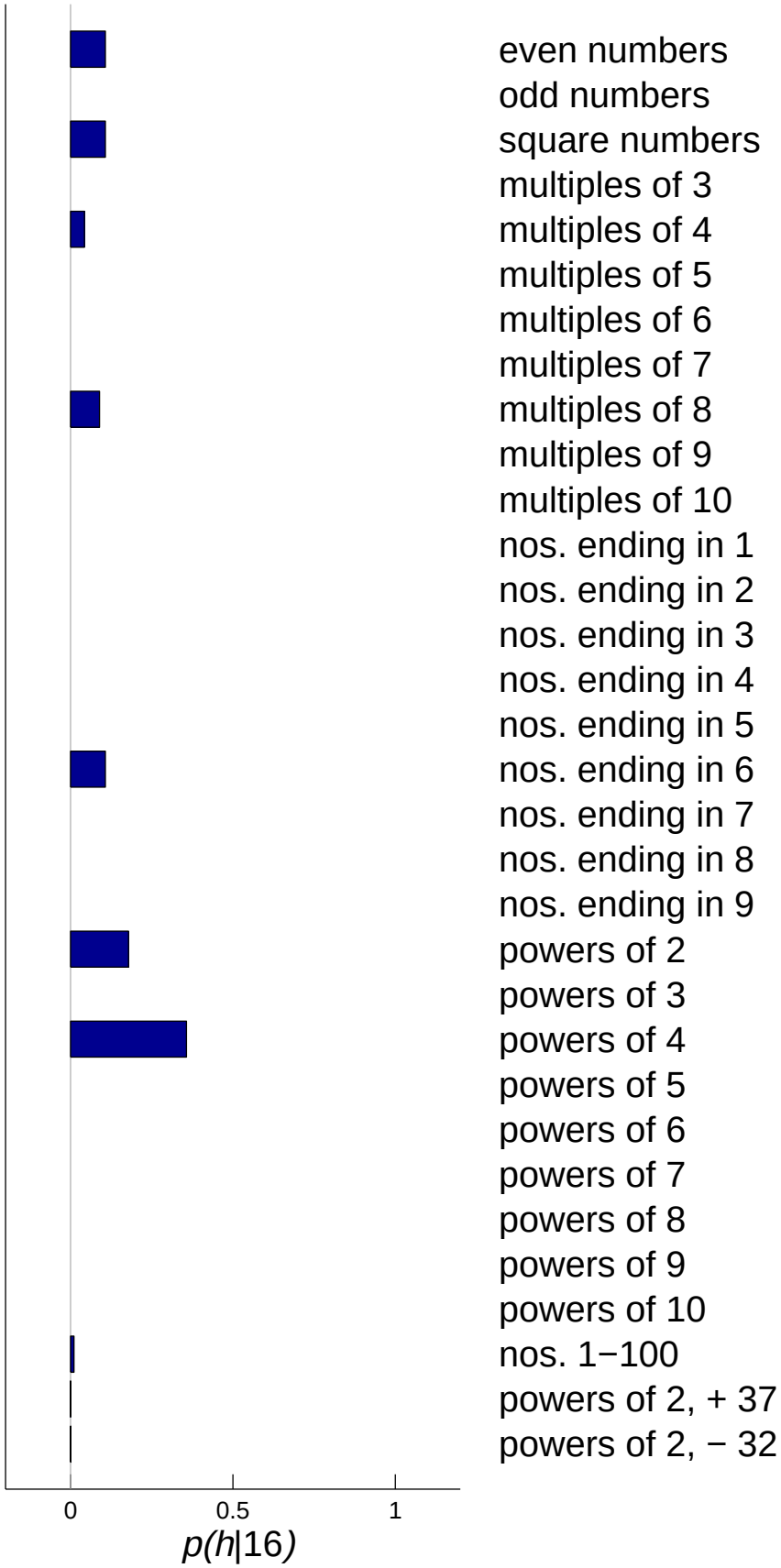

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0
0.5
1
p(h|16)

**Figure 2**

5000 hypotheses.) For their instructive value, we also consider a "default" hypothesis, corresponding to *all numbers between 1 and 100*, and two "unnatural" hypotheses, corresponding to *the powers of two except 32* and *the powers of two, and also 37*.

Observe that the distribution is rather "broad" or "flat": a number of hypotheses are assigned roughly similar probabilities, and all are rather small. [7] In general, broad probability distributions represent a high degree of uncertainty in our state of knowledge, whereas highly peaked distribution represent states close to certainty. The most peaked distribution possible has all of the probability mass on a single hypothesis; this corresponds to being absolutely certain that we know the extension of the concept. [8]

The broad distribution in Figure 2 is reasonable, because after seeing only one example of the concept, we are still rather unsure of how to generalize the concept to other numbers. The distribution is not completely flat, however; seeing the example 16 does give us *some* information about which numbers the program is likely to accept. Looking at Figure 2, certain hypotheses, like *powers of four*, *powers of two*, or *even numbers*, receive a relatively high probability. Others, like *powers of two except 32* or *numbers between 1 and 100*, receive a relatively low probability. Still others, like *odd numbers*, receive zero probability. Figure 3 shows the probability distribution $p(h|16, 8, 2, 64)$ after we have seen four examples. Now we are pretty certain that the program accepts all and only the powers of two, which is represented by a highly peaked distribution concentrated on that one hypothesis.

[7] Many roughly equal probabilities implies that they *must* all be much less than 1, due to the normalization condition.

[8] The degree of uncertainty in a probability distribution can be quantified by information-theoretic measures such as the entropy, $J(h|X) = -\sum_{h \in \mathcal{H}} p(h|X) \log p(h|X)$. See Cover & Thomas (1991) for an introduction to information theory, and Attneave (1959), Dretske (1981), and Bobick (1987) for some applications of these ideas in cognitive science.

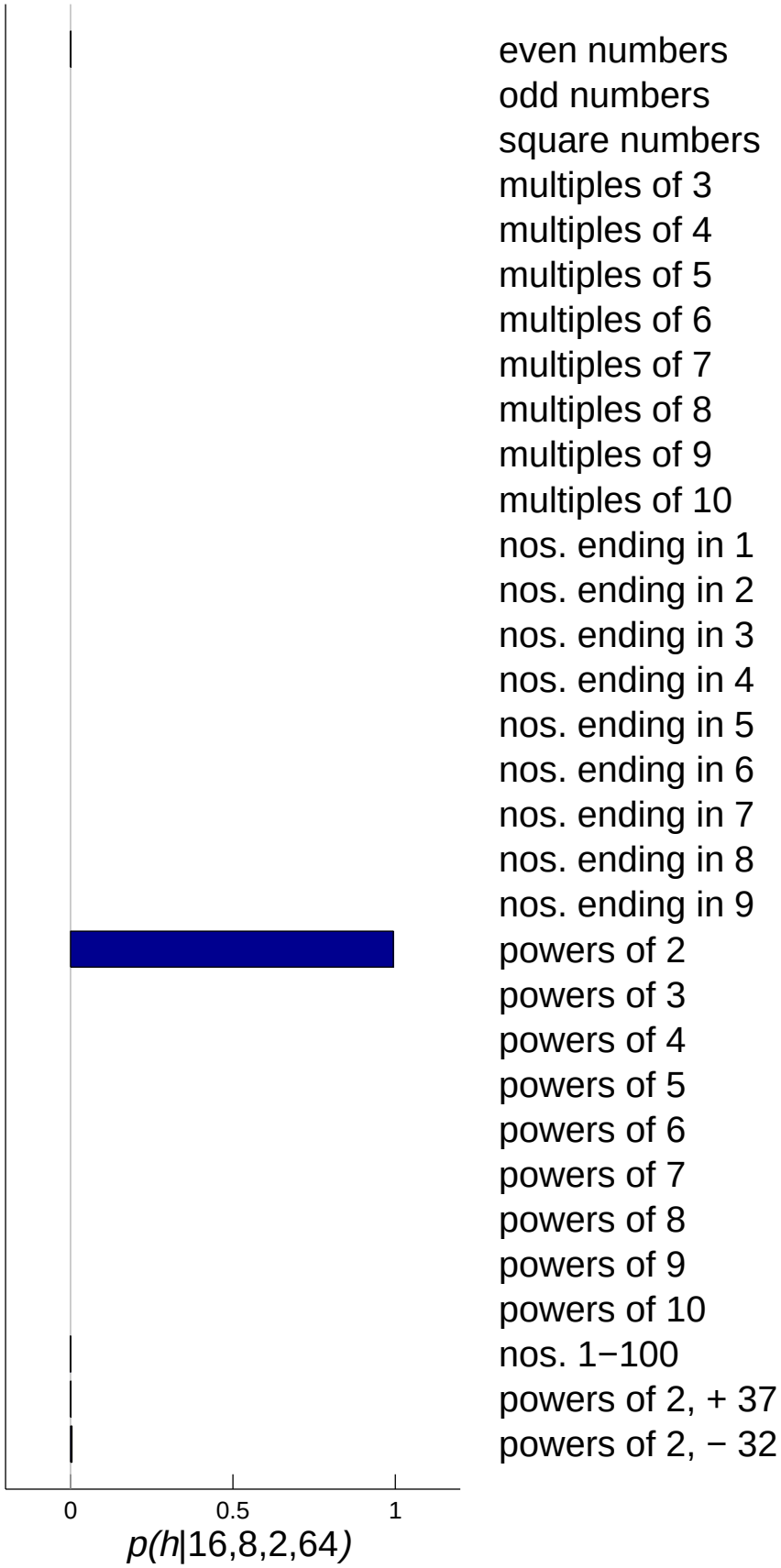

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0
0.5
1
p(h|16,8,2,64)

**Figure 3**

### 2.3.2 How is this knowledge acquired from observed examples?

Having shown how to represent our more or less uncertain knowledge about a concept's extension using probability distributions, we now consider how these probabilities could be determined from the observed examples. In general, the probability assignments $p(h|X)$ can be determined by Bayes' rule,

$$p(h|X) = \frac{p(X|h)p(h)}{p(X)}, \tag{2.1}$$

and thus depend on the product of the two terms $p(X|h)$ and $p(h)$. [9] The *likelihood* $p(X|h)$ measures the probability that we would observe the examples $X$ if $h$ were in fact the true extension of the concept. The *prior probability* $p(h)$ measures how probable we think it is that $h$ is the extension of the concept *before* we have observed any examples. The *posterior probability* $p(h|X)$ measures our belief in $h$ *after* we observed the examples $X$.

Bayes' rule is central to the theory of concept learning, because it capture the intuitive interaction between our prior beliefs about the extensions of possible concepts and our preference to avoid unexplained coincidences in the relation between concepts and their examples.

[9] There are two (linked) reasons why, for the moment, we don't need to worry about the denominator $p(X)$ in Equation 2.1, First, it is independent of $h$ and only serves to enforce the normalization constraint $\sum_{h \in H} p(h|X) = 1$. Second, because the hypotheses in $\mathcal{H}$ are assumed to be a mutually exclusive and exhaustive set of events, the laws of probability tell us that $p(X)$ can be expressed strictly in terms of the other two terms in Equation 2.1: $p(X) = \sum_{h \in \mathcal{H}} p(X|h)p(h)$. Notice how this expansion of $p(X)$ ensures the normalization constraint:

$$\begin{aligned}
\sum_{h \in \mathcal{H}} p(h|X) &= \sum_{h \in \mathcal{H}} \frac{p(X|h)p(h)}{p(X)} \\
&= \sum_{h \in \mathcal{H}} \frac{p(X|h)p(h)}{\sum_{h' \in \mathcal{H}} p(X|h')p(h')} \\
&= \frac{\sum_{h \in \mathcal{H}} p(X|h)p(h)}{\sum_{h' \in \mathcal{H}} p(X|h')p(h')} \\
&= 1.
\end{aligned}$$

### 2.3.3 The prior: $p(h)$

First consider the term $p(h)$. Now one might ask, how could we have any idea about the probability that some hypothesis $h$ is the true extension of the concept before we have seen any examples? However, as I argued above, we do in fact have such beliefs, and they embody our fundamental ideas about conceptual naturalness. In the number concept game, before we have seen even the first example 16, we do think it more likely that this program (or *any* program) picks out all and only the powers of two than that it picks out all the powers of two except 32, or all of the powers of two and also 37. A child learning the word "dog" knows that it is more likely that this word (or *any* word) picks out all and only the dogs in the world, than that it picks out all dogs except Aunt Sally's, or all dogs and also the Lone Ranger's horse (but nothing else). Not only are priors empirically real, they are in principle *necessary* for any kind of inductive inference to succeed (Goodman, 1955; Watanabe, 1985; Mitchell, 1980); recall my argument above that *all powers of two except 32* would be a better explanation of the observed examples 16, 8, 2, and 64 than *all powers of two* if not for the former's high *prima facie* implausibility.

Figure 4 illustrates a possible prior distribution over the simple number-game hypothesis space from Figure 2. The distribution is quite flat, reflecting our large degree of uncertainty before any examples have been observed. However, it crucially gives very low weight to bizarre hypotheses such as *all powers of two except 32*, placing essential constraints on the inductions we can make from any finite set of observations. Let's also assume that we are somewhat biased towards the two simplest hypotheses *even numbers* and *odd numbers*, reflected in those two hypotheses' relatively higher prior weight in Figure 4. This bias will make things a little more interesting later on when we compare alternative Bayesian models, but is otherwise inconsequential to the main points of this chapter.

Of course, the distribution depicted in Figure 4 represents only one possible state of knowledge that a learner could be in before observing any examples of the number concept. The appropriate priors – and the resulting generalization behavior – would

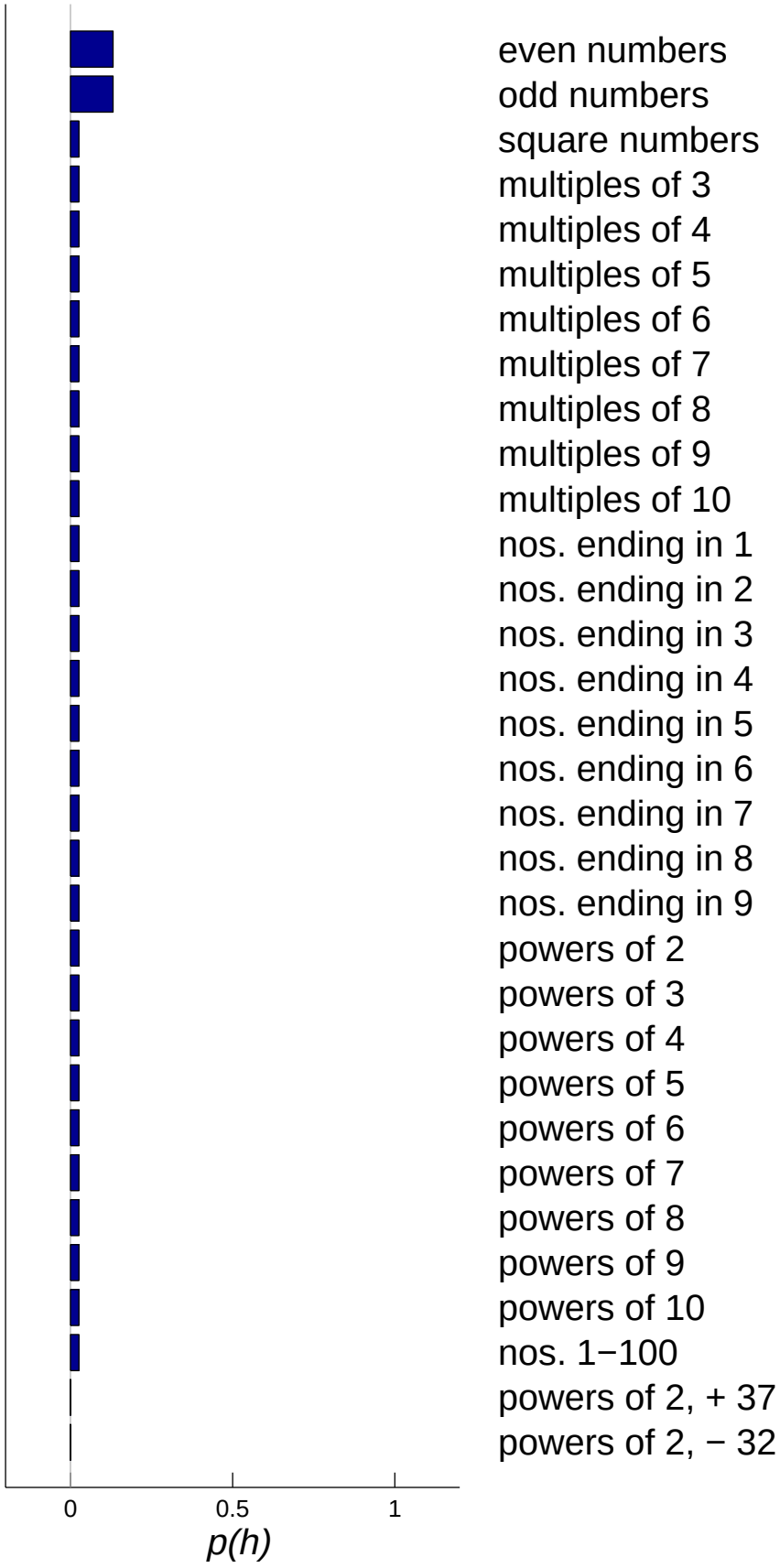

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0
0.5
1
p(h)

**Figure 4**

clearly be different in the case of a child who knows only how to count and pick out even and odd numbers, but nothing of powers, multiples, and other mathematically distinguished subsets of integers; or in the case of a mathematician who knows about primes, triangular numbers, Fibonacci numbers, and many other distinguished subsets.

More dramatically, a particular person might, in different contexts, have knowledge that leads him to generalize quite differently from exactly the same stimuli. Suppose we know that the numbers we are observing represent some physically meaningful quantity, such as cholesterol levels in the blood. Suppose moreover that the computer, instead of implementing some arbitrary program, is programmed to detect healthy blood levels of cholesterol (although of course we don't know what those healthy levels are – that's the learning challenge!). As before, we are given a few positive examples of the concept – numbers representing examples of healthy cholesterol levels – and asked to judge the probability that a new number also represents a healthy level. Given 1200, 1500, 900, and 1400 as examples of healthy cholesterol levels, 1183 seems quite likely also to be healthy while 400 seems rather unlikely to be healthy. However, given the same four numbers as positive examples in the number-program domain, 1183 would seem much less likely and 400 much more likely to be acceptable by the program.

Traditional models of concept learning have not been able to accomodate the range of background and contextual knowledge that human learners may have available to them (Heit, 1997), and this severely limits their applicability. In the Bayesian framework of this thesis, such knowledge effects can be modeled and understood as shifts of the learner's prior probability distribution over the hypothesis space of possible concept extensions, or as shifts of the hypothesis space itself. The different roles of prior knowledge in concept learning, and their effects on generalization behavior, will be explored extensively over the case studies of the next three chapters.

It should also be noted that there is no clear line between the choice of a hypothesis space and the choice of a prior. Omitting a particular hypothesis from the hypothesis space is equivalent to including it but assigning it a prior probability of zero. Hence

we can think of the choice of hypothesis space as an initial, qualitative prior which is then subject to further refinement in the assignment of $p(h)$.

### 2.3.4 The likelihood: $p(X|h)$

We now turn to the likelihood term $p(X|h)$. To compute the probability of observing the examples in $X$ given that hypothesis $h$ is the true extension of the concept, we require some assumption about the process that generates the examples and how it depends on the hypothetical extension. This is where the construct of a generative model, introduced above, enters into the formal theory.

A simple and powerful generative model that I will invoke frequently in this thesis is the *strong sampling* model, already alluded to above:

> **Strong sampling:** the observed examples of a concept are sampled randomly and independently from the concept's (unknown) extension.

I call this model "strong sampling" because it treats the examples as a random sample and makes the strong assumption that this sample comes from the true concept. Of course, strong sampling is only one possible generative model for the examples of a concept. Other important possibilities include:

- **Strong sampling with identification noise:** the examples are usually random and independent samples from the concept's extension, but with some small probability they are instances from outside the extension that have been incorrectly identified as positive instances.

- **Weak sampling:** the examples are generated by some random source independent of the true concept, and then labeled positive or negative according to their membership in the concept.

- **Feedback**: the observed examples are provided in the form of feedback occasioned by the learner's own use of the concept. Feedback may be either positive, *i.e.* confirmations of correctly identified instances, or negative, *i.e.* corrections of incorrectly indentified instances.

- **Helpful teacher**: the observed examples are deliberately chosen to be in some sense optimally informative about the concept's extension.

Each of these models suggests a somewhat different learning strategy, and there are natural situations under which each might be the most appropriate. Moreover, they can all be formalized in a Bayesian framework; I will have more to say about this later on. However, I have three principal reasons for taking strong sampling as the default. First, it is in some sense the most basic model; of the others listed, the first two are variations on it and the second two invoke resources (a deliberate teacher or feedback source) which may often not be available. Second, strong sampling describes in simplified form the most typical instances of supervised concept learning. Imagine that objects in the world appear according to some fixed (known) probability distribution $p$. Suppose also that every time a positive instance of concept $C$ appears, there is some (perhaps quite variable) probability $q$ that it will be labeled (*e.g.* in word, gesture, image) by a competent user of $C$ (*e.g.* by mommy, daddy, a friend, a child). Assuming $p$ and $q$ are independent (which is quite an assumption), strong sampling is the right generative model. Of course, word learning and other natural situations are never as simple as this or any of the other models we might write down in one or a few equations. Yet, it's somewhere to start. In Appendix C, I will consider some ways to make this model more realistic by incorporating the possibilities of mislabeled examples, homonyms and polysemy.

The final reason to focus on strong sampling is that it embodies a crucial insight into how people can learn concepts from positive examples only. In particular, it allows the likelihood term $p(X|h)$ in Equation 2.1 to implement our preference for avoiding unexplained coincidences in the relation between the concept to be learned and the examples we observe. To see why this is so, consider that the probability of randomly sampling any particular object out of a set containing $m$ objects is $1/m$. The probability of sampling any particular ordered pair of objects (under independent sampling and with replacement) from the same set is $1/m^2$, and in general, the probability of sampling any particular sequence of $n$ objects from the set is $1/m^n$. The key insight is this: as the size $m$ of the set we sample from gets *larger*, the

probability of obtaining any particular sequence of objects gets *smaller* by a factor that depends exponentially on the number $n$ of samples.

By the same reasoning, if $X$ denotes a sequence of $n$ randomly sampled examples, then the likelihood of observing this evidence given a particular hypothetical extension $h$ for the concept is simply

$$p(X|h) = \left[\frac{1}{\text{size}(h)}\right]^n, \tag{2.2}$$

if $h$ includes those $n$ objects, and 0 if it does not include one or more of them. We will often denote size($h$) as $|h|$. This simple equation will be crucial for understanding many of the results derived in the thesis, so it deserves a name: the *size principle*. In reading this thesis, if you remember only one equation other than Bayes' rule (Equation 2.1), let it be Equation 2.2 and the size principle.

The size principle can be seen as a quantitative form of Ockham's razor, "Entities should not be multiplied without necessity." Given the examples $\{x_1, \ldots, x_n\}$, Ockham prefers the hypothesis with the minimal "number of entities" necessary to explain their occurence. This is always just the set $X = \{x_1, \ldots, x_n\}$ itself if $X \in \mathcal{H}$, or else the smallest hypothesis in $\mathcal{H}$ containing the set $X$. Any larger hypothesis postulates "unnecessary entities", *i.e.* potential examples which have not actually been encountered. In general, the smaller the hypothesis, the fewer "unnecessary entities" it postulates and the more Ockham – and the size principle – like it. This idea of a thoroughly objective version of Ockham's razor, based on the range of data that a hypothesis can predict, is central in the Bayesian statistics literature (Jeffreys, 1961; Jefferys & Berger, 1992; Gull, 1989). Largely through the work of David McKay (1992), it has begun to penetrate into the statistical learning and neural computation fields. Recently, Muggleton (preprint) has introduced the idea in the context of inductive logic programming. This thesis is the first attempt to apply a size-based Ockham's razor to understanding the central problem of concept learning: how humans (or machines) can learn concepts from only one or a few positive examples.

To make the implications of the size principle more concrete, let us return to the number concept game, and suppose, as before, that we have seen one exam-

ple which the program accepts: $x_1 = 16$. Assuming that 16 is randomly sampled from all numbers less than 100 that the program accepts, then the likelihood $p(16|\textit{powers of two}) = 1/6 = .1666\ldots$, because there are 6 powers of two less than 100. Similarly, $p(16|\textit{even numbers}) = 1/50 = .02$, because there are 50 even numbers less than 100, $p(16|\textit{numbers less than 20}) = 1/20 = .05$, because there are 20 numbers less than 20, and so on. However, $p(16|\textit{odd numbers}) = p(16|\textit{powers of 10}) = 0$, because there is no way that 16 could have been sampled from either of those subsets. Figure 5 illustrates all the likelihoods $p(16|h)$ for our hypothesis space. The message is simple: while many hypothetical extensions are *consistent* with the one example, the likelihoods that they assign to it vary inversely with their size. Smaller, more specific hypotheses (*e.g. all powers of two*) assign a higher likelihood to – and thus receive greater inductive support from – a consistent example than do larger, more general hypotheses (*e.g. even numbers*).

Figure 6 shows how the likelihoods evolve as we see a few more examples. The differences between hypotheses already present after one example become increasingly exaggerated with each successive example, due to the size principle's exponential dependence on the number $n$ of examples observed (see Equation 2.2). After four examples have been observed, the difference in likelihood between *all powers of two* and *even numbers* comes to $1/6^4 = 7.7 \times 10^{-4}$ vs. $1/50^4 = 1.6 \times 10^{-7}$. Don't be deceived by the small absolute values; what counts is the ratio, which favors *all powers of two* by a factor of almost 5000:1. This quantifies the earlier intuition that it would be a very suspicious coincidence indeed to observe four examples which were all powers of two, if in fact the program accepted all even numbers. To sum up the effect of the size principle in words:

> **Size principle:** smaller hypotheses are more likely than larger hypotheses, and they become exponentially more likely as the number of consistent examples increases.

One aspect of Figure 6 may be disturbing: after four examples, the preferred hypothesis *all powers of two* does *not* have the highest likelihood. In fact, the "bizarre"

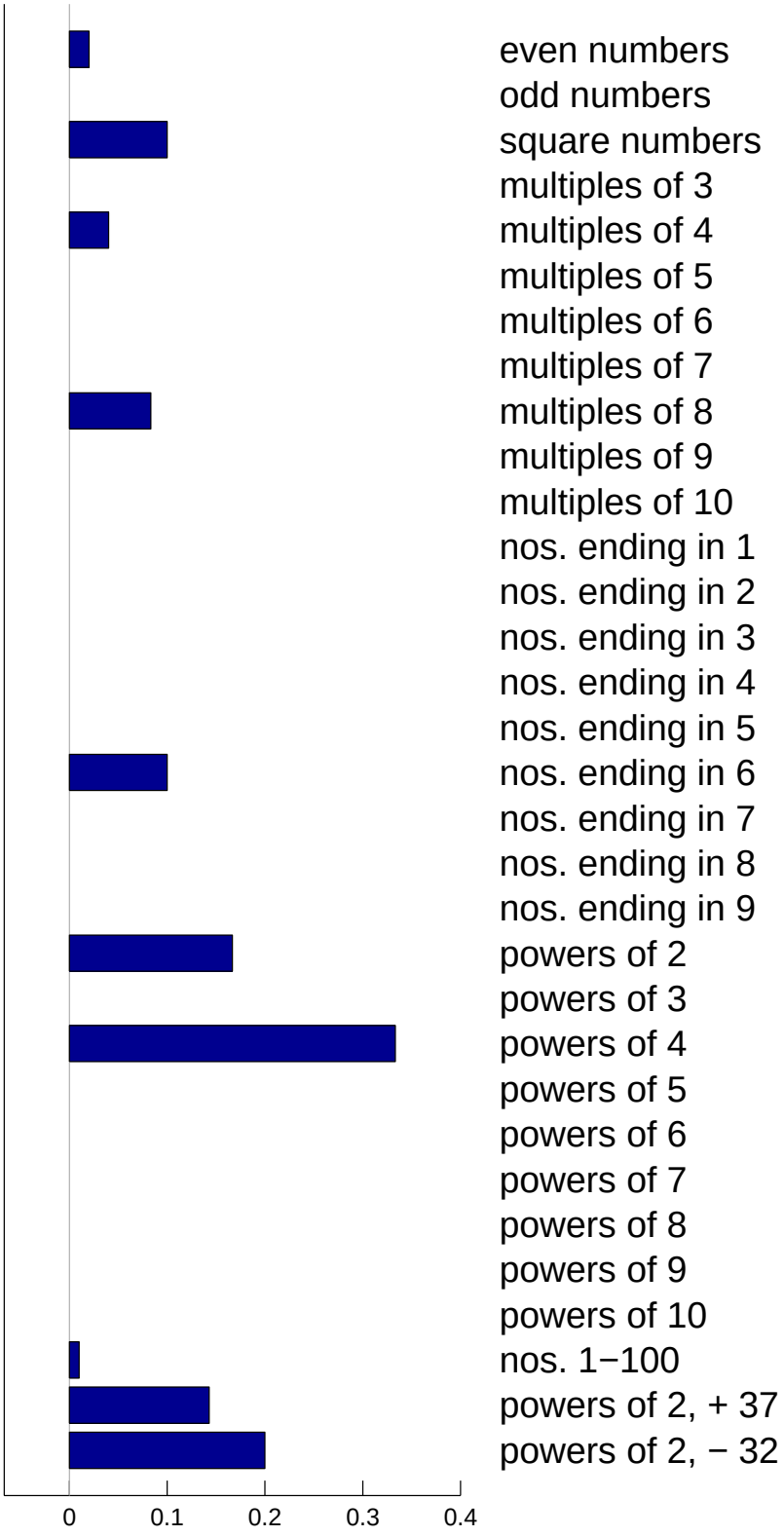

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0
0.1
0.2
0.3
0.4
p(16| h)

**Figure 5**

hypothesis *all powers of two except 32* beats out *all powers of two* by about 2:1 after four examples, because the former – missing 32 – is slightly smaller than the latter. (Recall the intuition from before that the non-appearance of 32 in the first four examples is a minor coincidence that is slightly better explained by this bizarre hypothesis.) This is why it is essential that not all logically possible hypotheses receive equal prior probability, and in particular, that such bizarre hypotheses receive near-negligible priors to compensate for their slight likelihood advantage.

Figure 7 shows how the combination of priors and likelihoods into posterior probabilities $p(h|X)$, given by Bayes rule (Equation 2.1), evolves from 0 to 4 examples. Notice that the single sharp peak in the posterior attained after 4 examples (Figure 7, right column) is present in neither priors (Figure 4) nor likelihoods (Figure 6, right column) alone. It is only the product of these two terms, balancing a pull towards our prior beliefs with a push away from unexplained coincidences, that shapes the "peak" of relatively certain knowledge which we can acquire from just a few positive examples.

### 2.3.5 How is this knowledge used to generalize the concept?

Finally, we consider how the learner can use this probabilistic knowledge about a concept's extension to judge which *new* stimuli will belong to the concept. Intuitively, when our knowledge takes the form of a distribution concentrated exclusively on a single hypothesis $h^*$, as is practically true after seeing the four examples 16, 8, 2, and 64 (Figure 7, bottom row), generalization should be a trivial matter. Any new stimulus $y$ either belongs to $h^*$ or not, and so the probability that the concept generalizes to $y$ is 1 if $y$ is in $h^*$ and 0 otherwise. In other words, when we have no remaining uncertainty about the concept's extension, generalization should be an all-or-nothing behavior.

However, our state of knowledge is frequently far from this ideal of certainty. In the case of a complex natural concept like "dog", merely observing a few examples will not be sufficient to determine its extension precisely. Even in the artificial world of the number concepttask, uncertainty is common. After we have seen only the one example

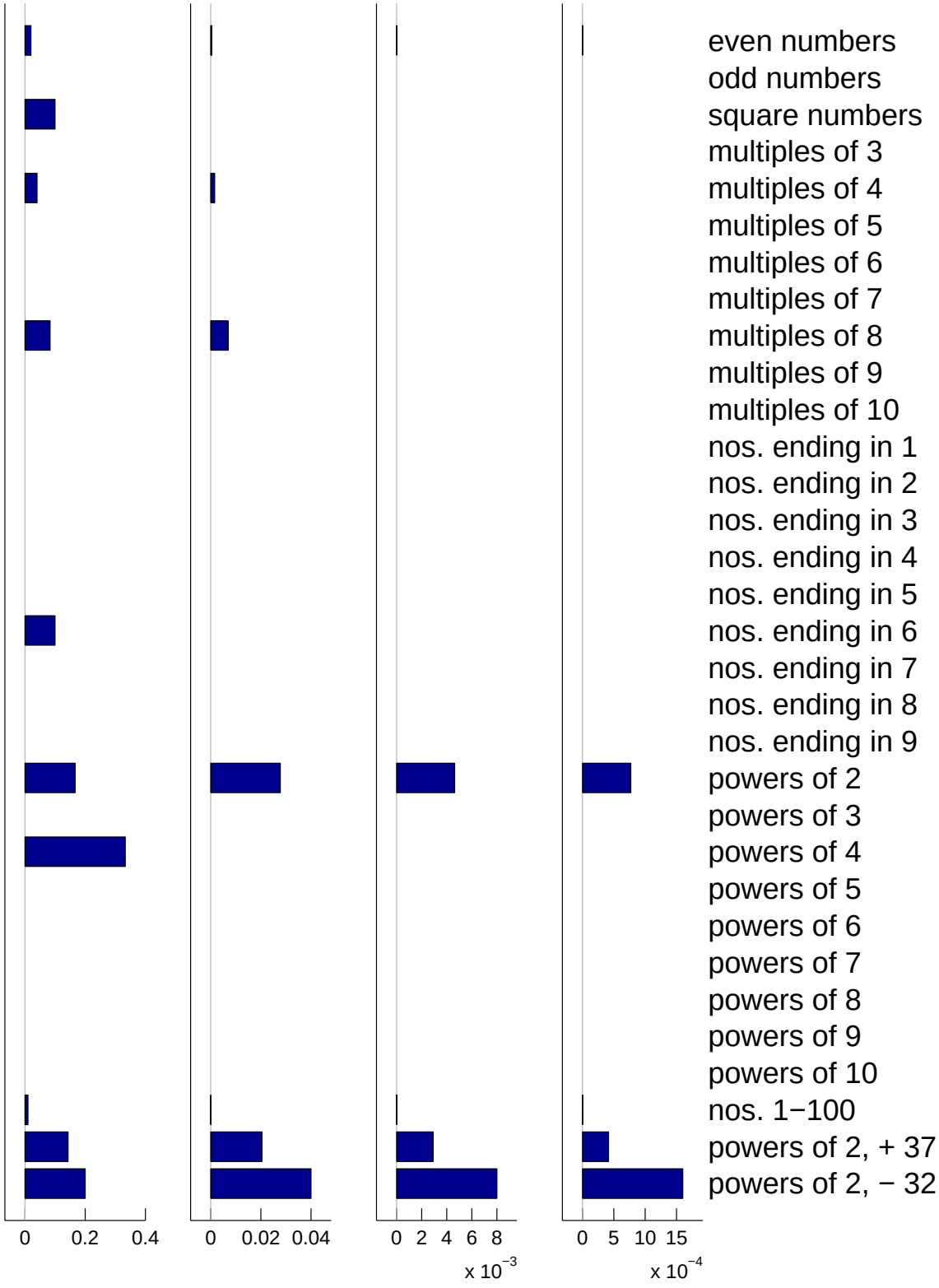

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0 0.2 0.4
0 0.02 0.04
0 2 4 6 8
x 10^−3
0 5 10 15
x 10^−4
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64| h)

**Figure 6**

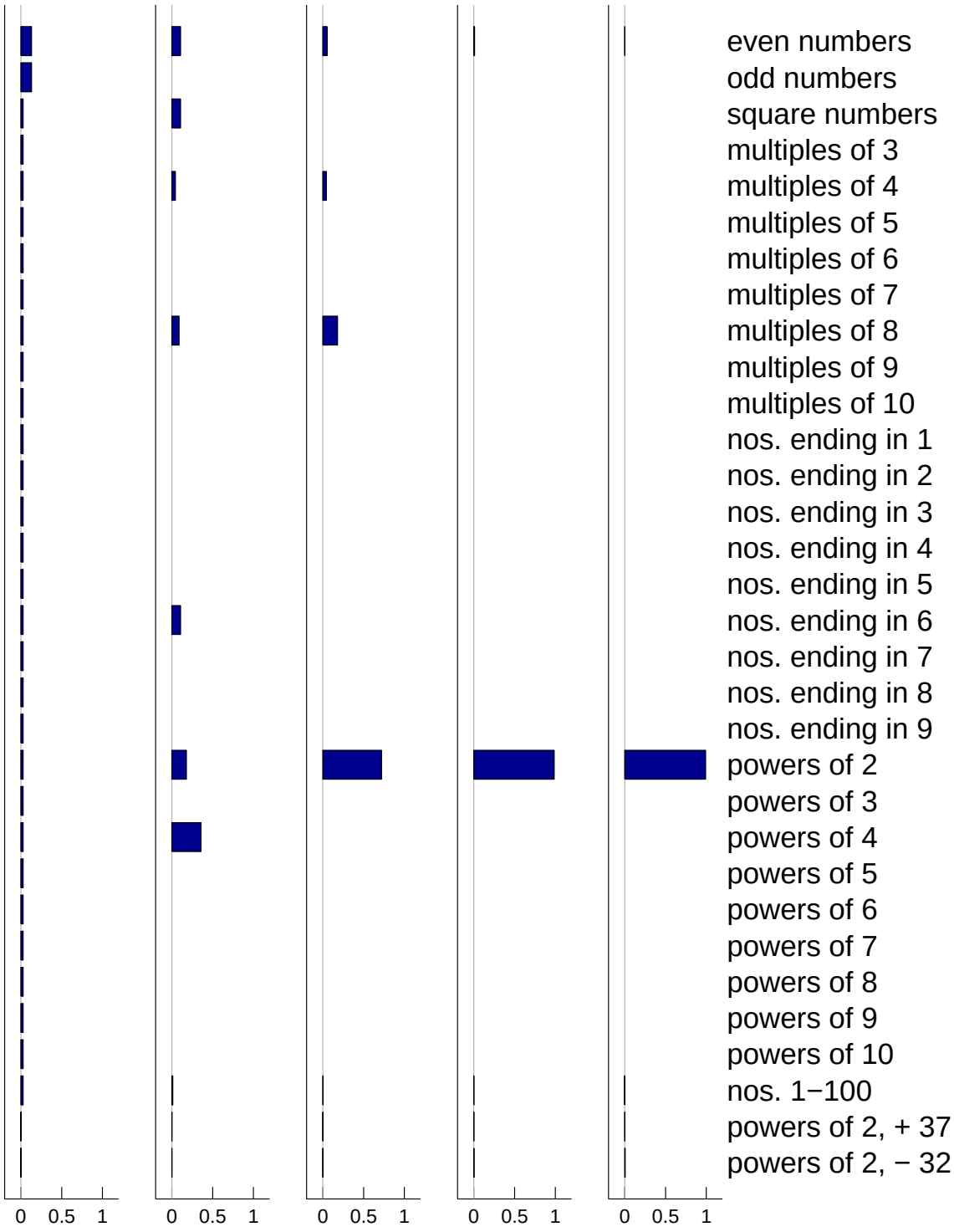

even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0 0.5 1
0 0.5 1
0 0.5 1
0 0.5 1
0 0.5 1
p(h)
p(h|16)
p(h|16,8)
p(h|16,8,2)
p(h|16,8,2,64)

**Figure 7**

16, our knowledge consists of a broad distribution over many possible extensions (Figure 7, 2nd column). After the second example, 8, our uncertainty has been reduced but there is still more than one reasonably plausible hypothesis (Figure 7, 3rd column). Only after some more examples, 2 and 64, does our knowledge approach a state of certainty (Figure 7, right column). Given such uncertain knowledge about the concept's extension as we have in Figure 7, columns 2 and 3, how are we to generalize, to decide which new stimuli – $17, 87, 4, 31, 32, \ldots$ – the concept might apply to?

In the Bayesian framework, we cast the problem of generalization as computing $p(y \in C|X)$, the probability that a new stimulus $y$ belongs to concept $C$, given the set $X$ of previously observed examples. Our hypothesis space $\mathcal{H}$ is what enables us to make the leap from the events we have seen, $X$, to the event we have not seen, $y \in C$. If we did accept a particular hypothesis $h$ as a concept's true extension, *i.e.* if our posterior probability was concentrated exclusively on $h$, then we should identify a new entity as an instance of the concept if and only if it falls inside $h$. Formally, $p(y \in C|X)$ would would equal 1 for all $y \in h$ and 0 for all $y \notin h$. When more than one hypothesis receives significant posterior probability, a Bayesian learner computes $p(y \in C|X)$ by *averaging* the predictions of all these hypotheses, weighted by their respective posterior probabilities $p(h|X)$. In other words, we "count up" the number of hypothetical extensions that include $y$ relative to the number of hypotheses that do *not* include $y$, with each hypothesis counted in proportion to its posterior $p(h|X)$. The resulting ratio gives the odds that $y$ does in fact fall under the concept.

Here's a concrete illustration. Suppose that after we have observed the examples $X$, there remain 10 consistent hypotheses. Suppose also, for simplicity's sake, that these 10 hypotheses receive equal posterior probability assignments. Then a new object $y$ which happens to fall inside 9 out of 10 of these hypotheses has 9:1 odds of being an instance of $C$, or $p(y \in C|X) = .9$. An object that falls inside 3 out of 10 hypotheses belongs to $C$ with probability .3, and so on. Note that each of the observed examples in $X$ is by definition consistent with all 10 of the candidate hypotheses, and thus belongs to $C$ with probability 1. Hypotheses with unequal

posterior probabilities present no difficulty; they are simply counted unequally, in proportion to their probability.

Formally, this prescription for generalization by averaging over the set of consistent hypotheses follows from the probability calculus and two additional premises: first, that whether or not $y$ belongs to $C$ is *conditionally independent* of the examples $X$, given that $h$ is the true extension of $C$; second, that $p(y \in C|h)$, the probability that $y$ falls under $C$ given that $h$ is the true extension of $C$, equals 1 if and only if $y \in h$ and 0 otherwise. From the conditional independence of $y \in C$ and X, we have [10]

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X). \tag{2.3}$$

Recall the version space $\mathcal{H}_y$, which denotes the subset of $\mathcal{H}$ that is consistent with the stimulus $y$. The condition $y \in h$ is then equivalent to the condition $h \in \mathcal{H}_y$. Because $p(y \in C|h) = 1$ if $h \in \mathcal{H}_y$ and 0 otherwise, the only terms that actually contribute to the sum over $h$ in Equation 2.3 are those with $h \in \mathcal{H}_y$. Thus we can rewrite the probability of generalization as

$$p(y \in C|X) = \sum_{h \in \mathcal{H}_y} p(h|X). \tag{2.4}$$

This is the formal statement of the above claim, that the probability of $y$ belonging to $C$ given the examples $X$ is computed by "counting up", or summing, the posterior probabilities $p(h|X)$ of all hypotheses including $y$.

The generalization probability from Equations 2.3 or 2.4 can be written in yet one more form, which is often the most useful expression for actually computing a

[10] The complete derivation of Equation 2.3:

$$\begin{aligned} p(y \in C|X) &= \sum_{h \in \mathcal{H}} p(y \in C, h|X) \\ &= \sum_{h \in \mathcal{H}} p(y \in C|h, X)p(h|X) \\ &= \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X) \end{aligned}$$

numerical answer or thinking about its significance: [11]

$$p(y \in C|X) = \frac{\sum_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n}{\sum_{h \in \mathcal{H}_X} p(h)/|h|^n}. \tag{2.5}$$

Here $\mathcal{H}_{X,y}$ denotes the version space of $X \cup \{y\}$ (*i.e.* the subset of hypotheses consistent with both $X$ and $y$). This form can be interpreted intuitively as follows. Based on the examples $X$, each hypothesis is assigned a score $p(h)/|h|^n$ that is the product of two terms: $p(h)$, which favors extensions with high a priori probability, and $p(X|h) = 1/|h|^n$, which favors smaller extensions to a degree that increases exponentially with the number of examples. The score thus formalizes the trade off between our prior beliefs about the natural extensions of concepts and the size principle, the statistical version of Ockham's razor favoring better explanations of the data (under the assumption of strong sampling). The probability of generalizing from the examples $X$ to the new stimulus $y$ is simply the ratio of the total score of hypotheses containing both $y$ and $X$ to the total score of all hypotheses containing $X$. The more hypotheses that include both $y$ and $X$, and the higher the scores those hypotheses receive, the more likely we are to conclude that $y$ belongs to the concept exemplified by $X$.

Figure 8 shows the probability of generalization for every number between 1 and 100, given the examples 16 (top row), $\{16, 8\}$ (middle row), and $\{16, 8, 2, 64\}$ (bottom

[11] This form comes from expanding the posterior $p(h|X)$ in terms of priors and likelihoods (Equation 2.1) and inserting the explicit value of the likelihood under the strong sampling model (Equation 2.2). From Bayes' theorem and the constraint that $\sum_{h \in \mathcal{H}} p(h|X) = 1$, we can write

$$p(h|X) = \frac{p(X|h)p(h)}{\sum_{h \in \mathcal{H}} p(X|h)p(h)}.$$

Plugging this into Equation 2.4, we obtain

$$p(y \in C|X) = \sum_{h \in H_y} \frac{p(X|h)p(h)}{\sum_{h \in \mathcal{H}} p(X|h)p(h)}.$$

Finally, we use the fact that $p(X|h) = 1/|h|^n$ if $h \in H_X$ (Equation 2.2), and 0 otherwise, to obtain

$$p(y \in C|X) = \frac{\sum_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n}{\sum_{h \in \mathcal{H}_X} p(h)/|h|^n}.$$

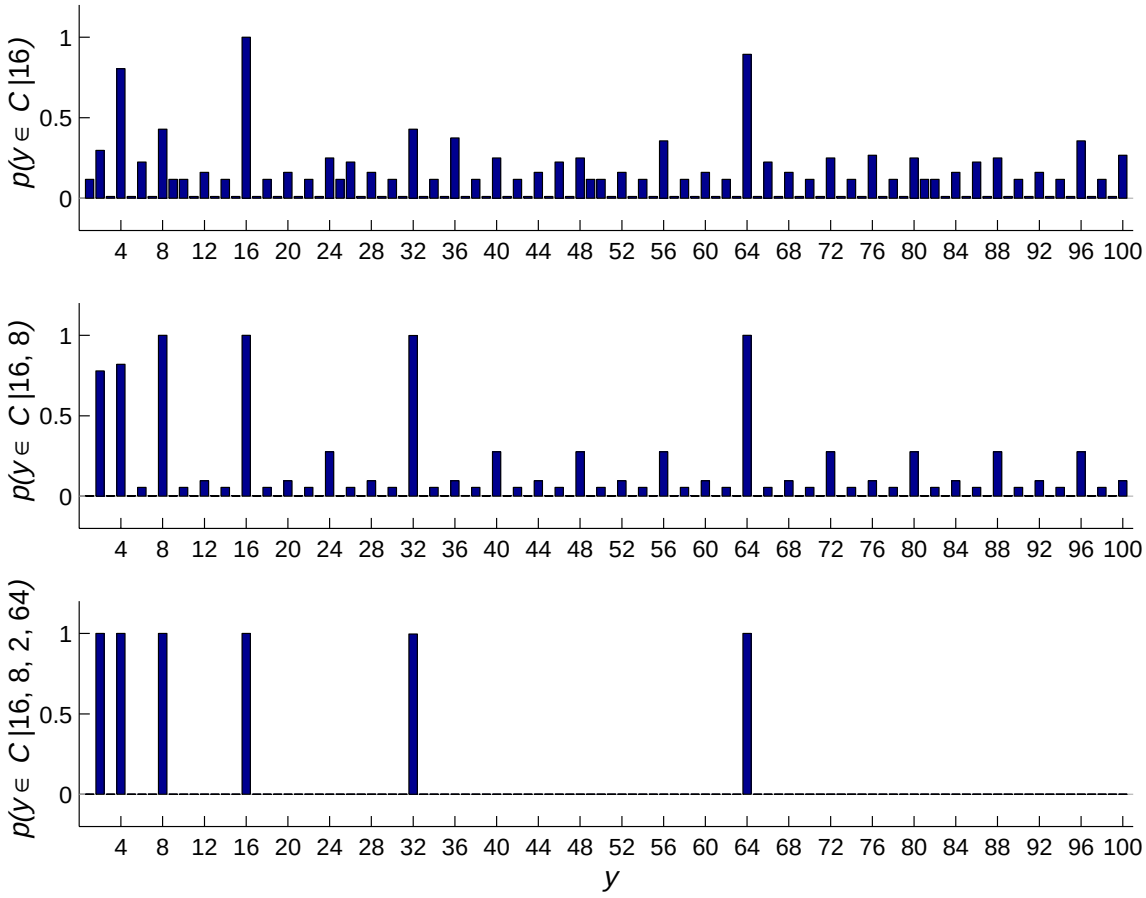



**Figure 8**

row). Observe the different kinds of generalization behavior that arise after one example and after four examples. Initially, given just 16, generalization is graded. Most numbers receive probabilities of belonging to the concept that are greater than 0 but significantly less than 1. Numbers very similar to 16, such as 4 or 64, receive higher probabilities than moderately similar numbers, such as 10 or 20, which in turn receive higher probabilities than very dissimilar numbers, such as 63 or 87. [12]

As more examples are observed, generalization rapidly sharpens up to a practically all-or-none function focused on the most specific hypothesis consistent with all the

[12]Notice how different these theoretical generalization gradients look from the real judgments of human subjects reported in Chapter 1, Figure 3. Remember this simulation uses only 30 hypotheses and leaves out many that subjects probably used in the actual experiment. In Chapter 5, we present a more comprehensive model with over 5000 hypotheses that captures human generalization behavior much better than the toy model considered in this chapter.

observations: *all powers of two*. Numbers that fit this rule receive probability close to 1, and all other numbers receive probability close to 0. Once again, the reason that *this* rule emerges – as opposed to other consistent hypotheses like *all even numbers* or *all numbers between 1 and 100* – and the reason it emerges so quickly – after just four examples – is the *size principle* (Equation 2.2): hypotheses corresponding to *smaller* extensions make the examples *more likely* than do hypotheses corresponding to larger extensions, and by a factor that increases exponentially with the number of examples observed.

Fundamentally, what determines whether generalization will be sharp or graded – rule-like or similarity-like – is the width (technically, the entropy) of the posterior probability distribution $p(h|X)$. The broader the posterior distribution, the less certain we are of the concept's true extension. When the probability mass is very spread out, as when we have seen only one example (Figure 7, 2nd column), many hypotheses each contribute to the average in Equation 2.3 and a broad gradient of generalization is the result. Figure 9 illustrates the process of averaging over all consistent hypotheses in this case. When $p(h|X)$ is concentrated narrowly on a single hypothesis, as when we have seen four examples (Figure 7, right column), only that one hypothesis contributes signficantly to the average in Equation 2.3 and generalization appears to be based in an all-or-none fashion on that rule. Figure 10 illustrates this concretely.

Thus we can see that the Bayesian framework has achieved the two main theoretical goals set out in Chapter 1. First, it explains how we can converge on a concept after seeing only a few positive examples. While many hypotheses may be equally consistent with the examples, the most specific hypothesis provides the *best explanation* for the occurence of the examples, in a statistical sense. Second, it accounts for how we can generalize when no single hypothesis dominates. By averaging the predictions of all consistent hypotheses weighted by their probability, we obtain a graded probability of generalization for any new object – a natural and principled measure of the similarity of that new object to the set of observed examples. At least in the domain of simple number concepts, we have succeeded in unifying rule-based and similarity-based generalization under a single theoretical framework. One basic

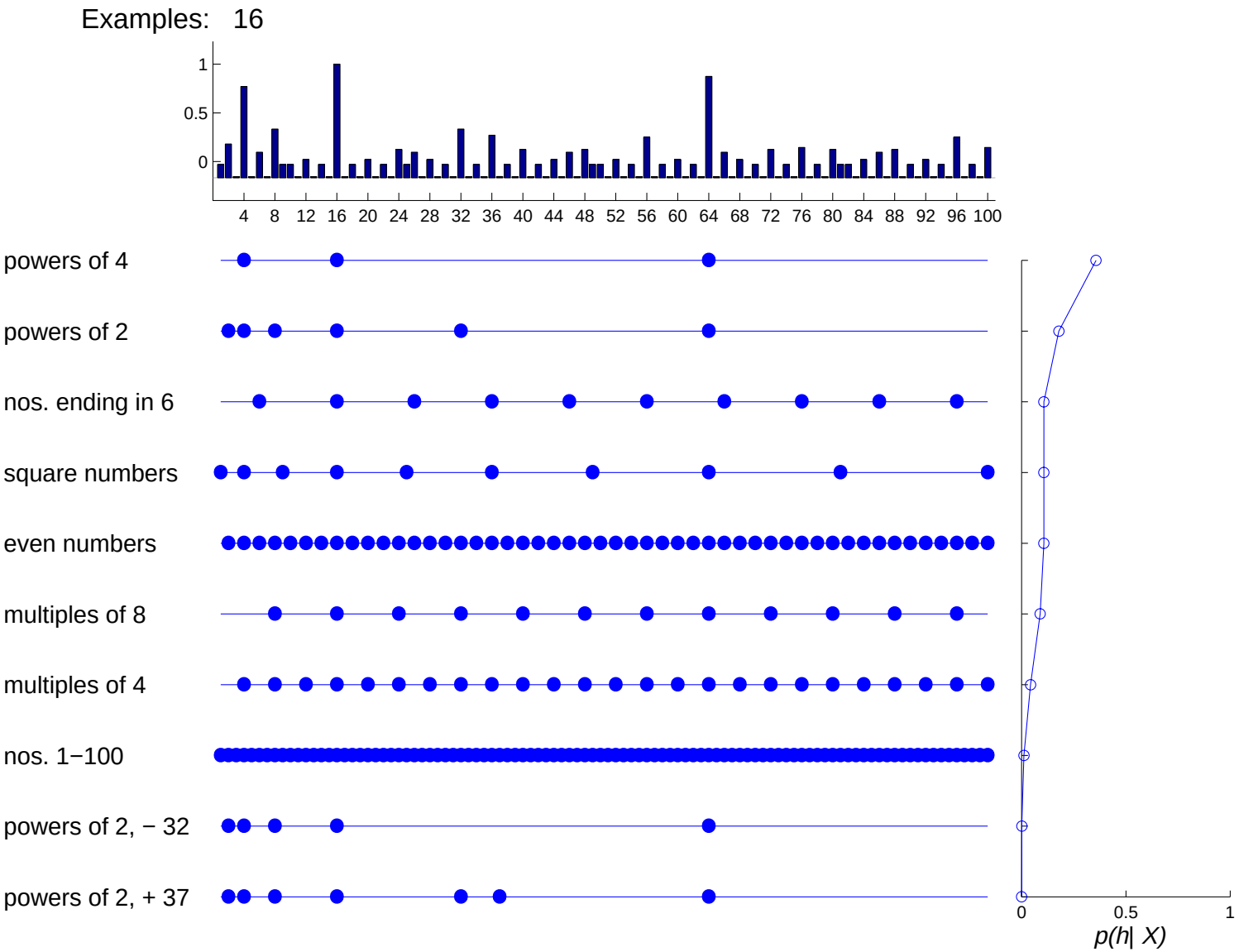

Examples: 16
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
powers of 4
powers of 2
nos. ending in 6
square numbers
even numbers
multiples of 8
multiples of 4
nos. 1−100
powers of 2, − 32
powers of 2, + 37
0
0.5
1
p(h| X)

Figure 9

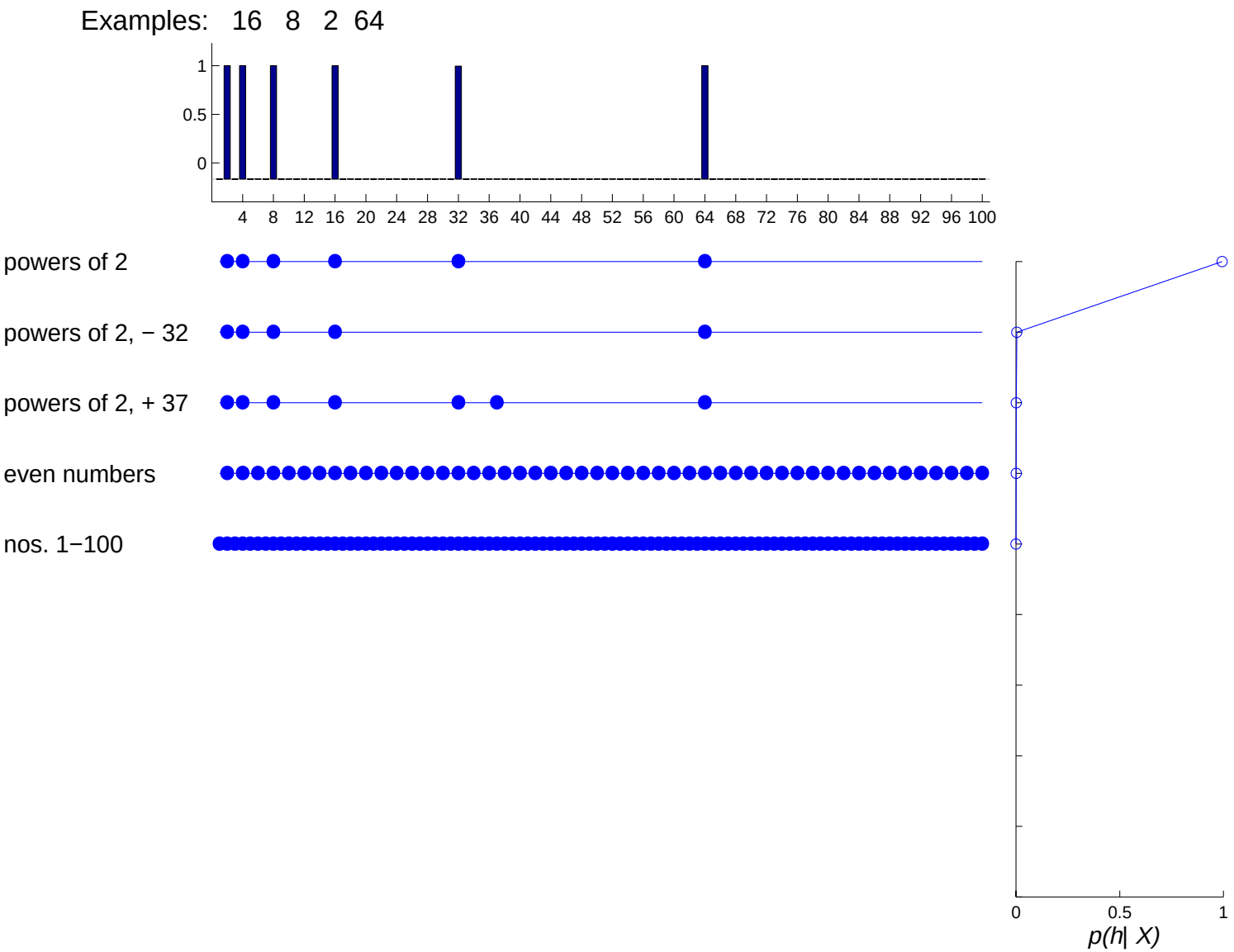

Examples: 16 8 2 64
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
powers of 2
powers of 2, − 32
powers of 2, + 37
even numbers
nos. 1−100
0
0.5
1
p(h| X)

Figure 10

principle, the *size principle*, explains how the observed examples can shape the relative strengths of competing rules, and more deeply, can shift the very character of generalization from what appears to be a similarity mechanism to what appears to be a rule-based mechanism.

## 2.4 Variants of the Bayesian framework

In this section, I will discuss a number of existing approaches to concept learning that can be seen as variants of the Bayesian framework developed here. Several of these were direct inspirations for the present work. Others correspond more or less to traditional rule- or similarity-based accounts of inductive inference. These correspondences tie into the theme of the last section; just as the Bayesian theory can explain both rule-like and similarity-like generalization behavior, it also contains as special cases the standard rule-based and similarity-based models for those behaviors.

I should note that this section is not meant to be an exhaustive review of the category learning literature. In particular, several important models of classification learning from cognitive psychology such as Kruschke's (1992) ALCOVE model, Anderson's (1991) "rational model", and various adaptive network models (*e.g.* Gluck & Bower, 1988; Estes, 1994) are not addressed here. These are models of discrimination learning, which require negative examples in order to generalize meaningfully and thus are not appropriate for the kinds of concept learning tasks that are the focus of this thesis. Appendix A discusses these alternatives and their difficulties in more detail.

To make clear the relationships between the different variants of Bayesian concept learning that we will consider, let me first summarize the four crucial ingredients of the Bayesian framework of this thesis:

**Summary of the Bayesian framework**

1. The notion of a constrained **hypothesis space** of possible extensions of a concept, and a probability distribution over that space, representing the learner's state of knowledge as to which entities a concept refers to;

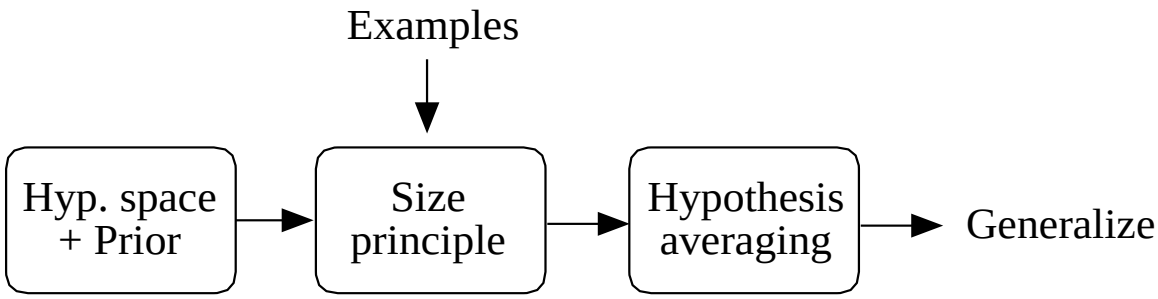



**Figure 11**

2. An informative **prior distribution** over the hypothesis space reflecting the background and contextual knowledge that the learner brings to the task;

3. The **size principle** for scoring the likelihood of hypotheses under the *strong sampling* generative model, favoring smaller consistent hypotheses with exponentially greater weight as the number of observed examples increases (Equation 2.2);

4. The notion of **hypothesis averaging**, *i.e.* integrating the predictions of multiple consistent hypotheses, weighted by their posterior probabilities, to arrive at the probability of generalizing a concept to a new entity (Equation 2.3).

Figure 11 illustrates the more or less sequential combination of these four ingredients that makes up the Bayesian framework. In the discussion to follow, I will refer to these ingredients both by number – ingredient 1, 2, 3, and 4 – or by their catch-phrases – **hypothesis space**, **prior**, **size principle**, and **hypothesis averaging**. What all the approaches below have in common is their use of a constrained hypothesis space (ingredient 1), plus some kind of Bayesian or statistical procedure for generalization, employing a combination of ingredients 2, 3, or 4. The approach I propose in this thesis is the first to combine all four ingredients. To distinguish it from the other alternatives, I will call this approach *Strong Bayes*, emphasizing the power that comes from combining Bayesian generalization with the strong sampling generative model.

After outlining each alternative approach below, I will evaluate its behavior on

the number concept game using the same toy hypothesis space we used above (minus the two "bizarre" hypotheses *all powers of two except 32* and *all powers of two, and also 37*). Besides giving a sense for the contributions of previous work, this will help us to see why all four ingredients – combined for the first time in this thesis – are necessary to understand how people learn and generalize concepts from just a few positive examples.

Also, I should make clear what the major alternative to each ingredient is. There is no alternative to ingredient 1; without some kind of constraints on the hypothesis space, generalization from finite evidence is hopeless. The alternative to ingredient 2, an informative prior distribution, is to assume that all hypotheses are equally likely a priori (*i.e.* $p(h) = 1/|\mathcal{H}|$, for all $h \in \mathcal{H}$). The alternative to ingredient 3, the size principle, is to use binary likelihoods that measure only whether a hypothesis is consistent with the data (*i.e.* $p(X|h) = 1$ if $h \in \mathcal{H}_X$, and 0 otherwise). The alternative to ingredient 4, hypothesis averaging, is to always pick the single most probable hypothesis (*i.e.* $h^* = \arg\max_h p(h|X)$) and generalize strictly according to its extension (*i.e.* $p(y \in C|X) = 1$ if $y \in h^*$, and 0 otherwise).

### 2.4.1 Simple hypothesis elimination: ingredient 1 only

I introduced the "simple hypothesis elimination" model in Chapter 1 as the most basic rule-based approach to induction, and I include this case here for completeness only. It assumes a constrained hypothesis space, but no statistical machinery in the form of ingredients 2-4. Hypotheses are not scored, but merely accepted or rejected based on whether or not they are consistent with the observed examples – the hypothetico-deductive method. This model is equivalent to Hovland's (1952) "communication analysis" of concept learning, and also forms the basis of Mitchell's (1982, 1997) CANDIDATE-ELIMINATION algorithm.

Without any way to compare hypotheses based on likelihood, this simple rule-based model has no way to generalize meaningfully from either the single example 16, or the four examples $\{16, 8, 2, 64\}$ (Figure 12). In both cases, multiple hypotheses are consistent with the observations and there is no principled reason (within this

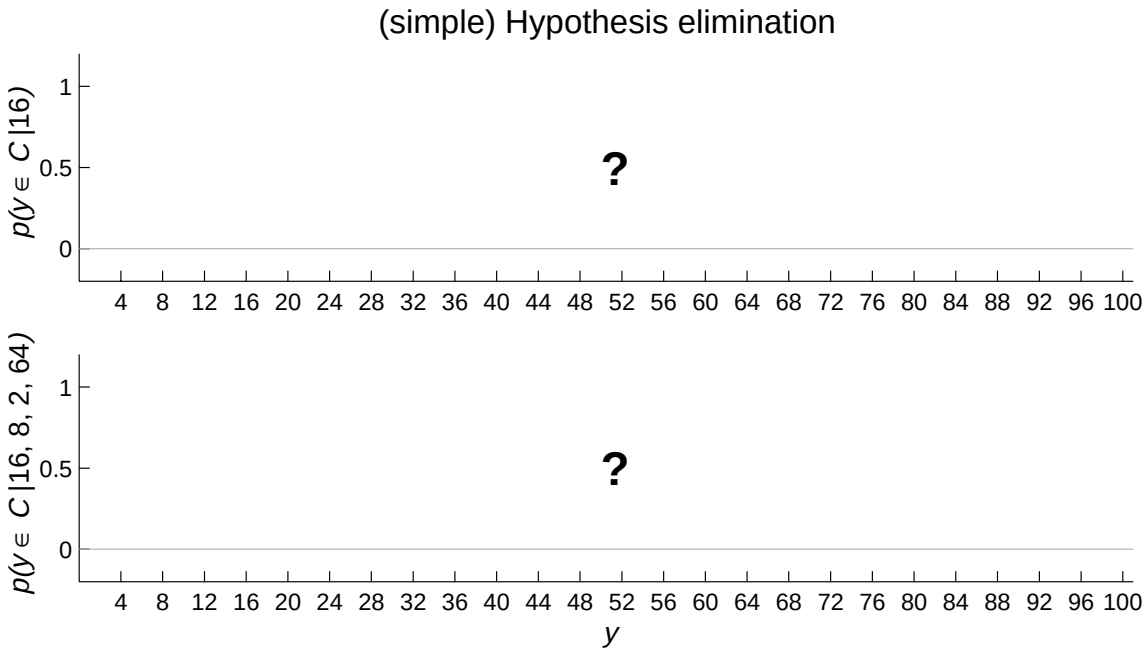



**Figure 12**

approach) to choose one hypothesis over the others. This was the main motivation in Chapter 1 for looking at more sophisticated rule-based models.

### 2.4.2 Ranked hypothesis elimination: ingredients 1 and 2

In Chapter 1, I suggested that the simple hypothesis elimination approach might be amended to ensure that a single best hypothesis could always be found. If the hypotheses are ranked according to some a priori measure of conceptual naturalness, we can always generalize by choosing the highest ranked hypothesis that has not been eliminated as inconsistent with the observed examples. This "ranked hypothesis elimination" approach is equivalent to ingredients 1 and 2 without ingredients 3 or 4. It also corresponds loosely to various proposals for concept learning and inductive inference in the philosophical literature, *e.g.* Fodor (1975), Howson & Urbach (1989).

Figure 13 shows the generalization behavior of this approach given the example sets $\{16\}$ and $\{16, 8, 2, 64\}$, assuming the prior distribution in Figure 4. Because this approach lacks ingredient 4, generalization is always based on a single rule and thus always appears all-or-none. Because it lacks ingredient 3, generalization will always

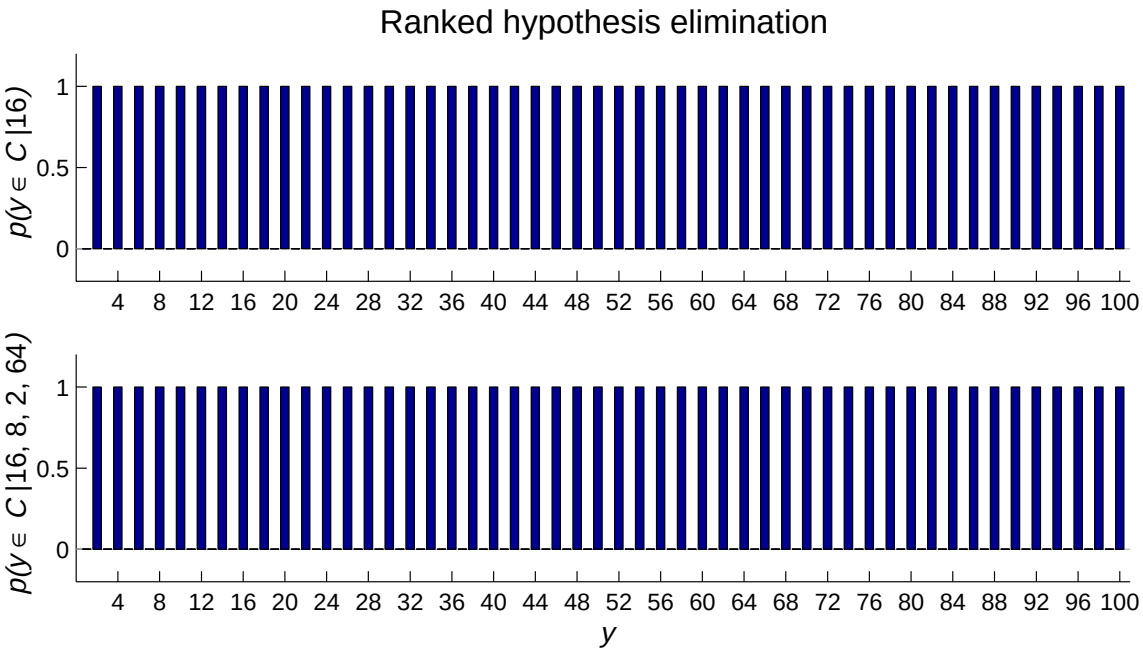



**Figure 13**

be determined strongly by the prior. The prior in Figure 4 places more weight on the hypothesis *all even numbers* than on other consistent hypotheses, and thus the learner generalizates from both $\{16\}$ and $\{16, 8, 2, 64\}$ to all even numbers with probability 1 (and all other numbers with probability 0).

### 2.4.3 MIN: ingredients 1 and 3

A more adaptive rule-based approach is the MIN algorithm for inductive inference. MIN says: always choose the smallest (*i.e.* most specific) hypothesis consistent with the observed positive examples. Because the smallest hypothesis also receives highest likelihood under the size principle, MIN can be thought of as implementing ingredient 3 without ingredients 2 or 4. Alternatively, MIN can be thought of as a *maximum likelihood* algorithm, a standard approximation to Bayesian inference which becomes equivalent to Bayes in the limit of infinite data, as the likelihood dominates the prior (Duda & Hart, 1973).

MIN has been proposed many times in many different inductive inference contexts. Popper (1959) suggests MIN as a way for scientists to choose among several competing

theories that all survive attempts at falsification by experiment. Bruner, Goodnow & Austin (1956), in the original cognitive studies of human concept learning, found evidence that MIN was a strategy commonly used by their subjects. In machine learning, Kearns & Vazirani (1994) and Mitchell (1997) present overviews of the theoretical properties of MIN algorithms. Valiant (1984) and Haussler (1988) proved some of the earliest formal learnability results for MIN, working in the probably approximately correct (PAC) learning paradigm (see Kearns & Vazirani (1994) for an introduction). In language acquisition, Pinker (1984), Berwick (1985) and Wexler & Manzini (1987) have all explored the usefulness of MIN for learning natural language grammars from only positive evidence. Last but not least, Feldman (1997) proposed a version of MIN in a framework for concept learning that was one of the direct inspirations for the present work.

With the exception of Feldman, none of these earlier authors thought of MIN in a probabilistic context, as a maximum likelihood algorithm or otherwise. Quite the contrary, Popper and many linguists working with MIN were actively opposed to the idea of statistical inductive learning. Instead, MIN has usually been justified in terms of some kind of asymptotic consistency argument: *i.e.* in the limit of infinite (or sufficiently many) data, it allows the learner to converge (or get arbitrarily close) to the true concept from only positive evidence. I'll return to the issue of justifying MIN in the discussion chapter.

Figure 14 shows MIN's performance on the number concepttask. Again, because MIN does not incorporate hypothesis averaging (ingredient 4), its generalization behavior is always strictly all-or-none. Given the one example 16, the minimal consistent hypothesis is *all powers of four*, so only 4 and 64 receive a nonzero probability of generalization. Given the four examples $\{16, 8, 2, 64\}$, the smallest consistent hypothesis is *all powers of two* and MIN generalizes exactly the same as Strong Bayes. In general, as more examples are observed, the extent of MIN's generalization can only become *broader* or remain unchanged. This conservatism is what gives MIN its asymptotic consistency guarantees; we never have to worry about "overshooting" the true concept. Because Bayesian inference asymptotically converges to the maximum

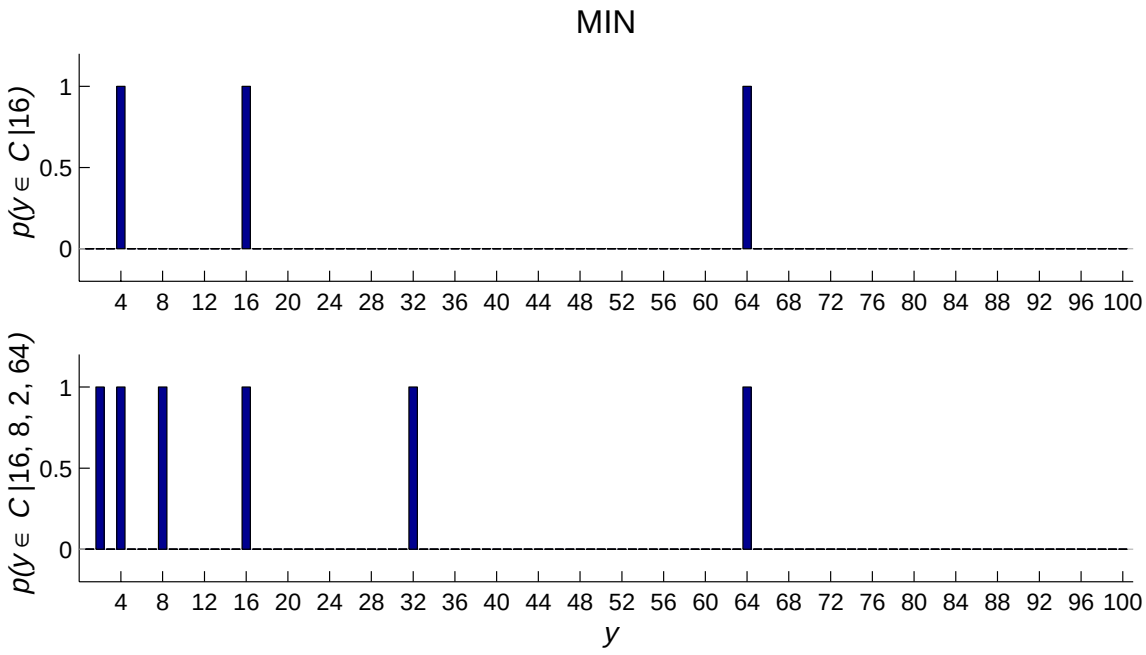



**Figure 14**

likelihood hypothesis, and MIN is a maximum likelihood algorithm under the strong sampling generative model, MIN and strong Bayes must become equivalent after sufficiently many examples have been observed, as happens here after four examples. But MIN's conservatism after just one example runs strongly counter to intuition. As MIN is such an important benchmark algorithm, I will return to it as a standard of comparison many times in subsequent chapters.

### 2.4.4 MAP: ingredients 1, 2, and 3

Even more sophisticated than MIN is the *MAP* algorithm. This approach is equivalent to adding an informative prior (ingredient 2) to MIN, or alternatively, to Strong Bayes, without the final step of hypothesis averaging (ingredient 4). That is, hypotheses (ingredient 1) are scored based on the combination of their a priori naturalness (ingredient 2) and the likelihood they assign to the data, under the assumption that the examples are random samples from the true concept to be learned (ingredient 3). The strong sampling generative model leads to the size principle favoring smaller hypotheses as the number of exampes increases, just as in Strong Bayes. However,

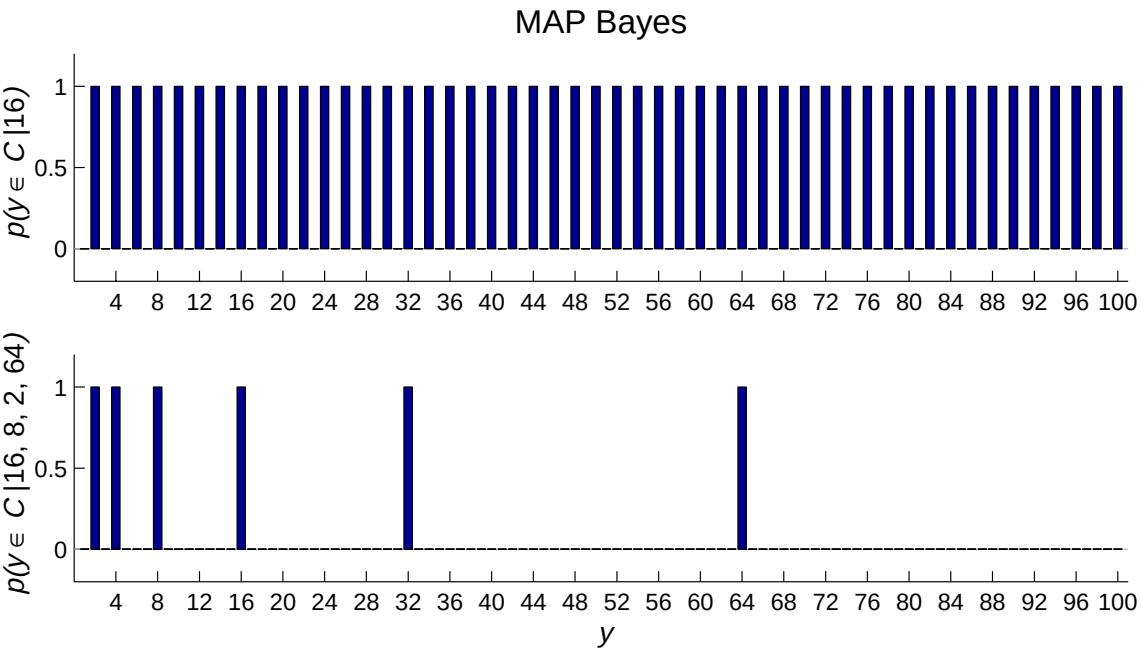



**Figure 15**

unlike strong Bayes, the MAP Bayes learner always chooses the single hypothesis with maximum posterior probability as the sole basis for generalizing to new stimuli (hence the name MAP, for "Maximum A Posteriori").

Figure 15 shows MAP Bayes generalization on the number conceptexample. In order to illustrate the difference between MAP Bayes and MIN (*i.e.* Maximum Likelihood), I've exaggerated the prior to favor the hypotheses *all even numbers* and *all odd numbers* over the other possibilities to an even greater extent than in Figure 4. (A priori, *all even numbers* and *all odd numbers* are now 10 times more likely than *powers of two* and all the rest.) As a result, after just the one example 16, the prior bias of 10:1 in favor of *all even numbers* outwieghs the likelihood's preference for smaller concepts like *multiples of four* (2:1 preference over *even numbers*) or *powers of 2* ($\tilde{9}$:1 preference); generalization extends to all even numbers with probability 1. In contrast, after 8, 2, and 64 have been observed, the likelihood term dominates the prior, with a preference of 5000:1 in favor of *powers of two* over *all even numbers*; generalization extends only to the powers of two. Thus, the strong sampling model (ingredient 3) allows MAP Bayes to converge on the most specific rule after just a few

examples. However, without hypothesis averaging (ingredient 4), MAP Bayes cannot account for similarity-based generalization, or for the shift from similarity-based to rule-based generalization as the number of examples increases.

A MAP Bayes model of inductive inference was first proposed by Watanabe (1960), and adapted to the case of concept learning by Hunt (1962). Horwich (1982) shows how a MAP Bayes model of induction illuminates a number of long-standing puzzles in the philosophy of science. Working in the context of inductive logic programming, Muggleton (preprint) has recently suggested that the principles of MAP Bayes could explain how people learn natural language grammars from only positive evidence. Muggleton's insight is essentially the same as mine here: a strong sampling-type generative model provides the key to learning from positive examples only, by trading off the size of a hypothesis against its a priori probability. His goal and theoretical analysis are quite different, however. Muggleton's focus is on establishing the asymptotic learnability (*i.e.* in the limit of infinite data) of languages and other complex logical systems, for which the size of each hypothesis in $\mathcal{H}$ may not be known in advance. A good part of the learning problem in Muggleton's case is estimating the size of the relevant hypotheses; hence, full Bayesian generalization by averaging over all hypotheses in $\mathcal{H}$ is not really an option for him. In contrast, my focus is on how people generalize concepts *for which they already have a notion of size* from just one or a few examples; hypothesis averaging is then essential, in order to understand the relation between similarity-based and rule-based modes of generalization.

### 2.4.5 Weak Bayes (uninformative prior): ingredients 1 and 4

We now turn to variants of the Bayesian framework which incorporate the idea of averaging the predictions of all consistent hypotheses (ingredient 4). Mitchell (1997) proposes a simplified version of optimal Bayesian concept learning, in which every hypothesis receives equal prior probability and every consistent hypothesis receives equal likelihood. The motivation for equal priors, or an "uninformative prior", comes from trying to make as few nonempirical assumptions as possible. The motivation for equal likelihoods comes from a weak sampling assumption, as opposed to the strong

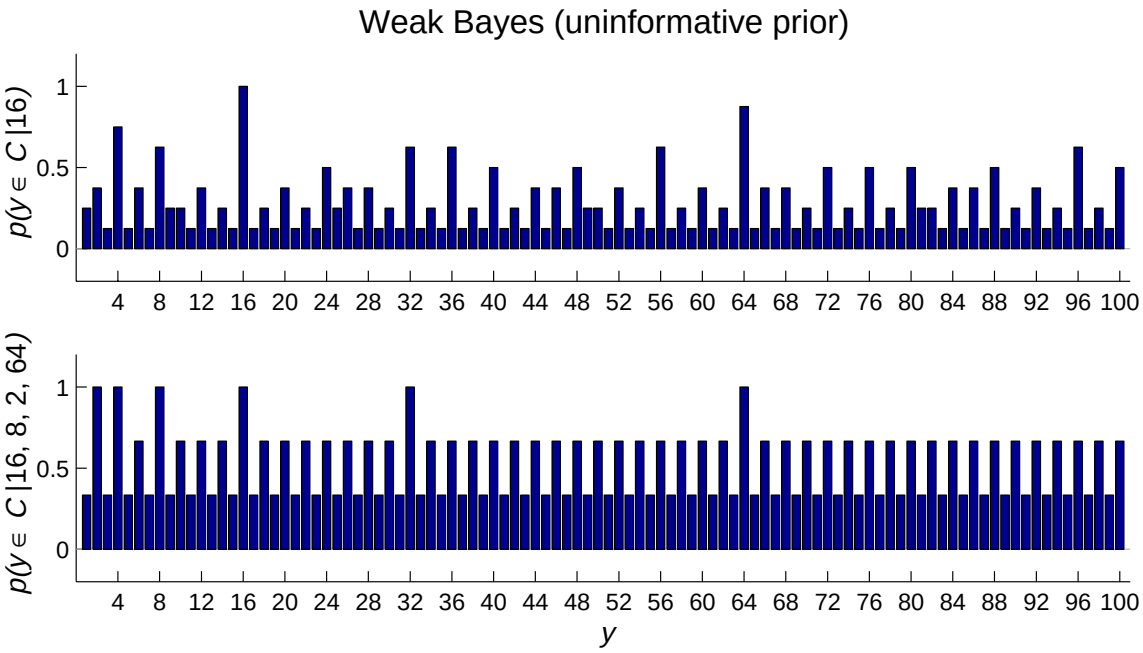



**Figure 16**

sampling assumption we make here which leads to the size principle. For this reason, I will call this approach *Weak Bayes*, and the approach developed in this thesis using the strong sampling model *Strong Bayes*.

To compute the probability of generalization, the predictions of all consistent hypotheses are averaged just as in Strong Bayes. Because of the uniform priors and binary likelihoods, the generalization probability takes a very simple form:

$$p(y \in C|X) = \frac{|\mathcal{H}_{X,y}|}{|\mathcal{H}_X|},$$

where $|\cdot|$ denotes the cardinality of the corresponding version spaces.

Figure 16 shows the behavior of this approach in the number concepttask. Unlike the previous two approaches, we now obtain a more intuitive *gradient* of generalization from the single example 16. However, after we observe the three more examples 8, 2, and 64, we do *not* converge to anything like all-or-none generalization on the powers of two. This is because the predictions of all consistent hypotheses are averaged independent of their size or the number of examples observed. In particular, there

are three hypotheses consistent with the set of examples $\{16, 8, 2, 64\}$: *all powers of two*, *all even numbers*, and *all numbers less than 100*. With no sense that one of these hypotheses should be more likely than the others, numbers such as 88 which fall under two consistent hypotheses receive a high probability (*i.e.* 2/3) of being accepted by the program, and numbers such as 87 which fall under one consistent hypothesis recieve a lower but still quite appreciable probability (*i.e.* 1/3). Intuition suggests both of these probabilities should be much closer to 0 than they are under this Weak Bayes model. Worse yet, it is *impossible* for Weak Bayes to converge on the true concept from only positive examples. No matter how many powers of two we see, we will never rule out any more hypotheses and the relative probabilities of generalization will stay constant at these values.

### 2.4.6 Weak Bayes (informative prior): ingredients 1, 2, and 4

The Weak Bayes approach becomes slightly more powerful when an informative prior (ingredient 2) is added – but only slightly. This version of Weak Bayes was thoroughly studied by Haussler, Kearns & Schapire (1994), who also introduced a stochastic variant they called the *Gibbs* algorithm that was inspired by work in the statistical mechanics of learning from examples (Seung, Sompolinsky, & Tishby, 1992). In cognitive psychology, Heit (1999) has presented a Bayesian analysis of category-based induction that is equivalent to this Weak Bayes model.

Formally, the posterior probability of a hypothesis is now given simply by renormalizing the prior over all those hypotheses consistent with the observed examples:

$$p(h|X) = \frac{p(h)}{\sum_{h \in \mathcal{H}_X} p(h)}. \tag{2.6}$$

Inserting these values for $p(h|X)$ into Equation 2.5, the probability of generalization becomes:

$$p(y \in C|X) = \frac{\sum_{h \in \mathcal{H}_{X,y}} p(h)}{\sum_{h \in \mathcal{H}_X} p(h)}. \tag{2.7}$$

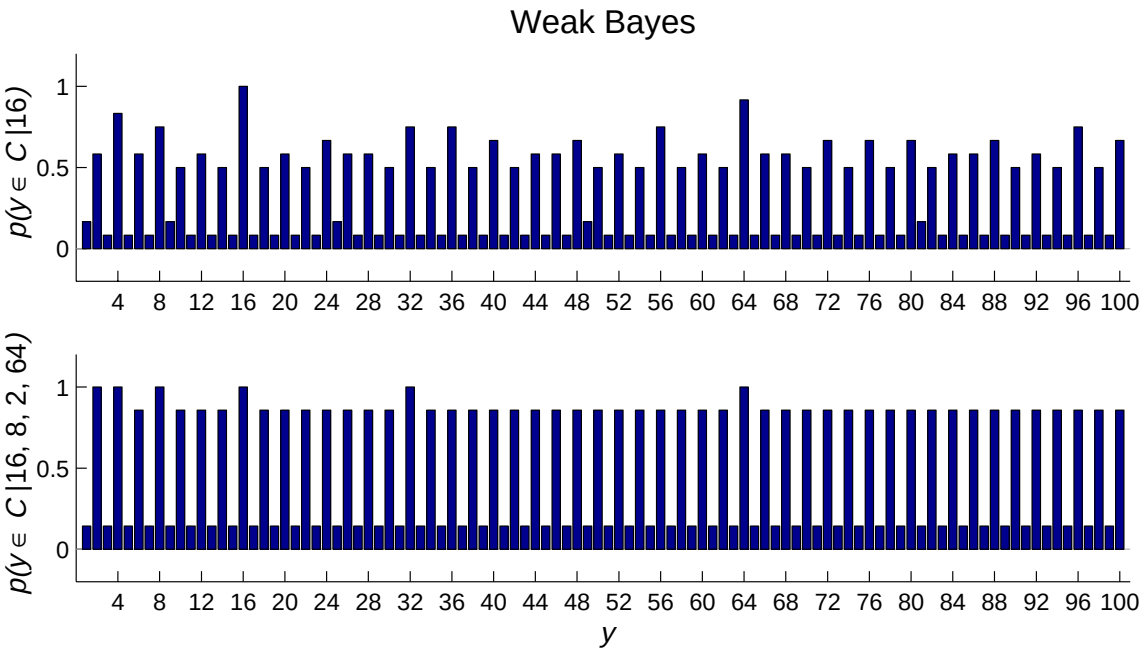



**Figure 17**

Figure 17 shows the behavior of this approach in the number concepttask. Because the prior is so weak in this case, there is little difference over the results without an informative prior (Figure 16). Generalization still follows a gradient similarity to the observed examples, with no sign of ever converging to a rule (due to the lack of ingredient 3). The one noticeable change from Figure 16 is that whether or not a number is even receives much higher weight in determining generalization, because of the strong a priori bias towards *all even numbers* in the prior (Figure 4).

From the standpoint of cognitive psychology, Weak Bayes models are important because they correspond to standard similarity-based approaches to concept learning. Recall the identification of hypotheses with "features" I suggested in Chapter 1 (*i.e.* two objects share a feature corresponding to each hypothesis that they both belong to). Then Weak Bayes with an uninformative prior (Equation 2.4.5) corresponds to computing similarity based on a simple count of features common to all examples in $X$ and the new object $y$, relative to the number of features common to all examples in $X$. Using an informative prior (Equation 2.7) corresponds to taking weighted versions of the same feature counts, with the weight of each feature determined by

the prior probability of the corresponding hypothesis. These models are thus very closely related in spirit to Tversky's (1977) classic feature-based models of similarity.

Shepard's (1987) derivation of the universal law of generalization (see discussion in Chapter 1 (Section 1.5.1) and Chapter 3 (Section 3.2.3)) can also be seen as a Weak Bayes model of concept learning under an informative prior. Shepard was trying to understand the rational basis of the similarity gradients that are observed empirically when organisms are required to generalize from a single example. He showed that under an appropriate Weak Bayes learning model, generalization gradients from a single stimulus approximate an exponentially decaying function of *distance* in the psychological space of stimuli. For readers not familiar with Shepard's work, we will go through a very similar derivation in Chapter 3, where we consider the problem of generalization in a continuous stimulus space, which was the concrete focus of his work.

Note that Shepard (1987) did not formulate his analysis in the explicit form of Weak Bayes; because he considered generalization from only a single example, he never faced the problem of nonconvergence illustrated in Figures 16 and 17. His work was one of the main inspirations for this thesis, because he was the first in psychology to give a rational analysis of generalization, and the first to show that a graded sense of similarity was the rational consequence of averaging over many consistent discrete hypotheses. Russell (1986) gives an alternative rational analysis for the universal law of generalization, focusing on stimuli represented by discrete features (see also Gluck, 1991). Tversky's (1977) theory, of course, also derives similarity from averaging over discrete features, but without the basis in a rational analysis of the inductive inference problem the organism is trying to solve. Stern (1991) explicitly noted that Tversky's model could be given a rational basis, using essentially a Weak Bayes argument. I will return to the connection between the Shepard and Tversky approaches to similarity when I discuss heuristics for concept learing (Chapter 5).

### 2.4.7 Conclusion

We have shown that many previous models of concept learning can be seen as limiting cases of the Bayesian framework presented in this thesis. In particular, different cases reduce to some of the most well-known learning algorithms based on either rules or similarity. MIN corresponds to a popular rule-based approach that is guaranteed to converge to the true extension of a concept given enough sufficiently many examples. Weak Bayes encompasses two of the major paradigms for modeling similarity-based generalization, Tversky's (1977) feature contrast theory and Shepard's (1987) "psychological space" analysis. We'll see more of both MIN and Weak Bayes in coming chapters, when we try to evaluate the extent to which rules or similarity are capable of describing the generalization behavior of human learners in real learning tasks.

In addition to showing the connections between my work and many previous models of concept learning, I hope this section has clarified the role that each ingredient of the Bayesian framework plays in guiding generalization. The constrained hypothesis space (ingredient 1) allows for the very possibility of generalization from finite evidence and is common to all approaches discussed above. The prior (ingredient 2) allows pre-existing conceptual biases (*e.g.* towards more familiar concepts like *all even numbers*) to override statistical information when very little data (*e.g.* 1 example) is available. The size principle (ingredient 3), a statistical principle derived from the strong sampling assumption, enables the learner to converge on the true concept after just a few examples. Finally, by averaging the predictions of all hypotheses weighted by their probability (ingredient 4), the learner can automatically shift between similarity-based and rule-based modes of generalization as a function of how confident he is of the concept's true extension.

It is really the combination of ingredients 3 and 4 – the size principle and hypothesis averaging – that allows us to model the interaction of rules and similarity in concept learning. Previous models that incorporate the size principle but not hypothesis averaging (*i.e.* MIN and MAP) can converge to the best rule after several examples, but do not show the phenomena of graded generalization by similarity af-

ter only a single example (Figures 14 and 15). Models that incorporate hypothesis averaging but not the size principle (*i.e.* Weak Bayes approaches) can generalize by similarity but can never converge to the true concept no matter how many examples are observed (Figures 13 and 14). This thesis is the first work to recognize the importance of both the size principle and hypothesis averaging for learning concepts from just a few positive examples, and thus the first theory capable of unifying both rule- and similarity-based generalization under a single explanatory framework.

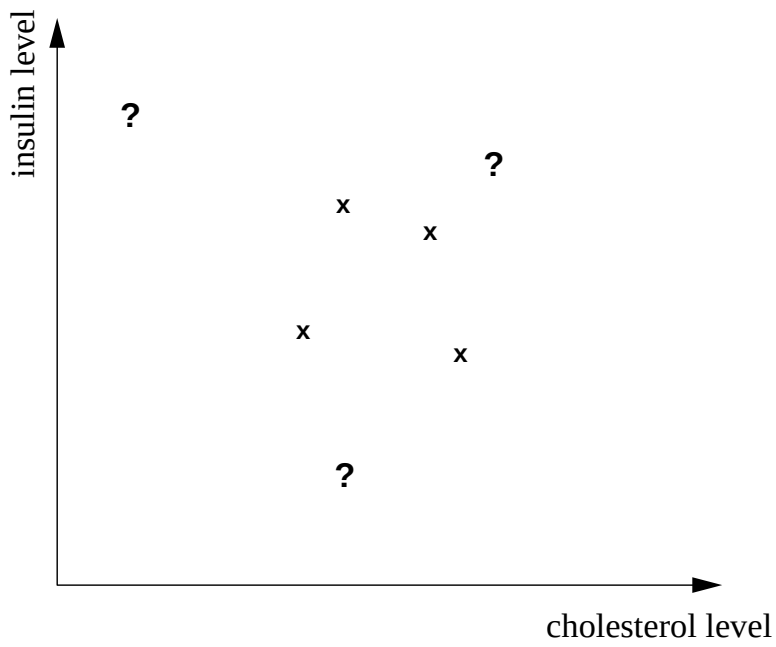
# Chapter 3

# Case Study #1: Learning concepts in continuous feature spaces

The first two chapters have laid out the computational problems of learning and generalizing concepts from just a few positive examples and shown how these problems can be addressed in a Bayesian framework. In the next three chapters, I present a series of detailed case studies of the framework, each in a different domain of concept learning. These case studies serve both an empirical purpose and a theoretical one. On the empirical side, they show how the Bayesian approach is able to predict actual human behavior on a range of different learning tasks and across different stimulus domains. On the theoretical side, these studies seek to explain what it is about these different domains – or really, about the different kinds of hypothesis spaces that people bring to (or generate in) these domains – that give rise to what looks like such different patterns of generalization.

## 3.1 Introduction

Let me introduce the domain for the first case study. Consider a population of people coming in for a medical checkup. Suppose that for each person, we can measure only two numerical features: their blood concentration levels of insulin and cholesterol. Thus we can represent each person by a point in a two-dimensional feature space




"healthy levels"

**Figure 1**

(Figure 1). Suppose that a doctor shows us four examples of people considered to have healthy blood levels of cholesterol and insulin (indicated by +'s in Figure 1), and our job is to judge whether people with different blood levels of these substances (indicated by ?'s in Figure 1) should also be identified as healthy.

Although it may not look like it at first, this "healthy levels" task is a very simple instance of concept learning. The concept "healthy levels" corresponds to some subset of the possible pairs of blood concentration levels of insulin and cholesterol – some region of the feature space depicted in Figure 1. We are given four random positive examples from this concept, four points known to belong to this region, and we are required to generalize the concept, *i.e.* to decide which other points are likely to belong to this region. Obviously, this is a very impoverished kind of concept learning. Nonetheless, like the number concept game intoduced in the last chapter, it is worth our attention because it isolates the essential challenge of inferring how far and in what ways to generalize a concept beyond the observed examples, in a form that is

analytically tractable and very amenable to empirical study in human subjects. Compared to the last chapter's number concept game, the prior knowledge relevant for this task is relatively simple: cholesterol and insulin are both important biochemical substances produced naturally by the human body in some range of healthy concentrations, becoming unhealthy when they exceed some unknown maximum healthy level or fall below some unknown minimum healthy level. Moreover, the healthy levels task is very similar to other learning tasks studied empirically in cognitive psychology (Shepard, 1964; Nosofsky, 1986) and explored theoretically in the literature on pattern recognition (Duda & Hart, 1973) and machine learning (Mitchell, 1997). This will facilitate comparisons of the present framework with the results of previous work.

This chapter begins by reviewing how existing theoretical frameworks for concept learning could be applied to the healthy levels task. I will show that standard rule-based (*i.e.* MIN) and similarity-based (*i.e.* Weak Bayes) learning algorithms are each inadequate on their own, but can be thought of as special cases of a Strong Bayes model that yields intuitively plausible patterns of generalization from one or more positive examples. I will then present an experiment with human subjects on the healthy levels learning scenario. The Bayesian framework provides a close quantitative fit to how people actually generalize in this task, suggesting an explanation for some aspects of the human ability to generalize concepts from only a few positive examples.

## 3.2 Theoretical analysis

### 3.2.1 Classical approaches

Let us begin by considering how standard approaches from the rule- and similarity-based traditions behave on this task. For rule-based concept learning, we first need to specify a hypothesis space of possible rules for picking out the concept's extension; for similarity-based learning, we need to specify a similarity metric on this feature space.

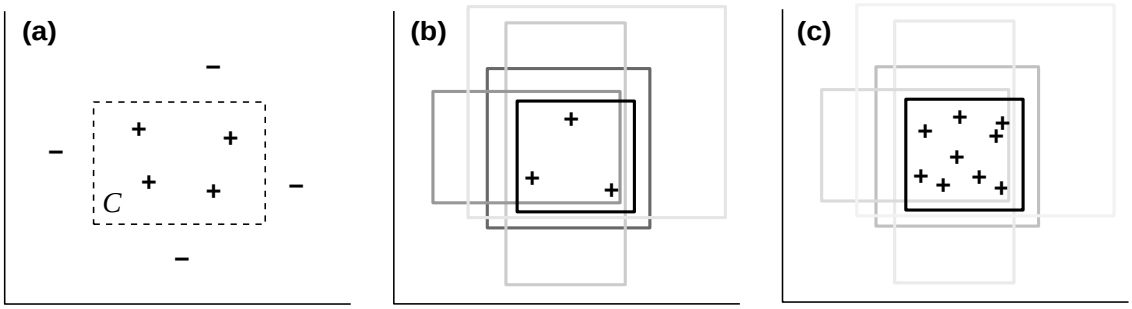



**Figure 2**

### Hypothesis space: axis-parallel rectangles

Suppose that the concept "healthy levels" applies to any individual whose cholesterol and insulin levels are each greater than some minimum healthy level and less than some maximum healthy level. Then the concept "healthy levels" corresponds to a *rectangle* in the two-dimensional feature space defined by cholesterol and insulin levels; in particular, a rectangle parallel to the coordinate axes of this space. If the extent of the concept along each of these dimensions is assumed to be independent a priori, then the hypothesis space $\mathcal{H}$ should consist of *all* axis-parallel rectangles in this space. Figure 2 illustrates the notion of a rectangle concept (a) and several possible hypotheses about its extension that are consistent with three examples (b).

In machine learning, the problem of learning axis-parallel rectangles is a common textbook example used to illustrate models of concept learning (Mitchell, 1997; Kearns & Vazirani, 1994), and it is also the focus of state-of-the-art theoretical work and applications (Dietterich et al., 1997). Rectangle learning tasks are not well known *per se* in cognitive psychology, but many experimental studies of human learning have investigated very similar tasks (Shepard, 1964; Goldstone, 1994; Nosofsky, 1986), with stimuli defined over two perceptually separable dimensions (*e.g.* size and brightness) and concepts defined by simple dimensional rules (*e.g.* size $\geq 3$, $2 \leq$ brightness $\leq 5$). Shepard's (1987) analysis of generalization in a two-dimensional separable feature space, one of the direct inspirations for this case study, also uses a hypothesis space

of axis-parallel rectangles. [1]

### Similarity metric: city-block distance

We can assume that we have a metric, or distance measure, along each dimension separately, corresponding to the difference between two levels of cholesterol or insulin measured in the relevant units. Let $d_j = d_j(x, y)$ denote the difference between two points $x$ and $y$ on dimension $j$. What we need is a way to combine these one-dimensional distances into a two-dimensional distance, and then a way of turning that 2-D distance into a measure of similarity suitable for generalization.

The natural way to combine distances along two separable dimensions is called the *city-block*, or $\mathcal{L}^\infty$, metric (Shepard, 1987). In the city-block metric, the distance between two points is just the sum of their distances along each dimension individually. In order to allow for different units on each dimension, or for the possibility that one dimension is more important than the other in determining similarity, we will allow for two scale parameters $\sigma_1$ and $\sigma_2$ which weight the relative contribution of each dimension to the total distance (analogously to "north-south" vs. "east-west" block length). Thus we have

$$d(x, y) = \frac{d_1}{\sigma_1} + \frac{d_2}{\sigma_2}. \tag{3.1}$$

As a result of several classic experimental studies (Shepard, 1987; Nososfky, 1992), it is generally accepted that similarity in separable feature spaces is best modeled as an *exponentially decaying* function of city-block distance:

$$\text{SIM}(y \rightarrow x) \quad = \quad \exp\{-d(x, y)\}$$

[1] The case of generalization in multidimensional feature spaces with integral dimensions is not treated in this thesis, but there are several possibilities ways it might be treated. One possible hypothesis space suggested by Shepard (1987) includes only regions with equal extent along both dimensions, *i.e.* square instead of rectangular regions. Another possible hypothesis space includes rectangles with all possible orientations in feature space, *i.e.* not only those aligned with a partcular set of axes in the space. This second possibility is discussed in Appendix C. Both kinds of hypothesis spaces would give rise to a Euclidean metric for generalization from a single example, but would give rise to very different behavior after multiple examples are observed. I leave it to future work to decide between these and other possibilities.

$$= \exp\{-(\frac{d_1}{\sigma_1} + \frac{d_2}{\sigma_2})\}.$$

### 3.2.2 Generalization based on rules vs. similarity

Neither a hypothesis space of possible rules nor a pairwise similarity metric is a complete model of concept learning. Any rule-based algorithm must specify some way of choosing the *best* hypothesis to guide generalization; any similarity-based algorithm must specify how an object's pairwise similarity to each of the examples gives rise to its *setwise* similarity to the set of examples as a whole. (Recall the extensive discussion of these issues in Chapter 1.) Here we consider two simple algorithms, MIN RULE and MAX SIM, chosen as representatives of the wide range of possible approaches from the rule- and similarity-based traditions. Both of these algorithms have a history in human and machine learning, and both will soon be recognizable as variants of the Bayesian framework that were discussed at the end of Chapter 2.

Given a set of examples $X = \{x_1, \ldots, x_n\}$, the MIN RULE algorithm chooses the most specific rule $h_{min}$ (*i.e.* the rule with the minimal extension) that is consistent with all the examples in $X$. Under the city-block metric above, $h_{min}$ is just the rectangle with smallest volume that contains all the positives. [2] The probability of generalization is then 1 for any object $y \in h_{min}$, and 0 for any object outside $h_{min}$. In the machine literature, MIN RULE is the standard algorithm for concept learning with axis-parallel rectangle hypotheses (Haussler, 1988; Kearns & Vazirani, 1994; Mitchell, 1997). In cognitive psychology, MIN RULE was one of the principal strategies adopted by human concept learners in the classic studies of Bruner et al. (1956), and was also the main focus of Feldman's (1997) studies of perceptual concept learning – another direct inspiration for the present work.

The MAX SIM algorithm defines the probability of generalization to a new object

[2] More generally, the minimal rule can be defined without a metric, as long we have a partial ordering of hypotheses by inclusion (Mitchell, 1982; Feldman, 1997).

$y$ as the maximum pairwise similarity of $y$ to each of the examples in $X$,

$$p(y \in C|X) = \max_i \text{SIM}(y \rightarrow x_i). \tag{3.2}$$

Under the exponentially decaying similarity function introduced above (Equation 3.2), MAX SIM is equivalent to

$$p(y \in C|X) = \exp\{-\min_i(d_1^i/\sigma_1 + d_2^i/\sigma_2)\}, \tag{3.3}$$

where $i$ ranges over all examples in $X$ and $d_j^i$ denotes the distance from $y$ to example $i$ along dimension $j$. Because it is much easier to analyze, we will consider a variation on Equation 3.3 that assumes that the generalization function is separable in the two dimensions of our feature space:

$$p(y \in C|X) = \exp\{-(\min_i d_1^i/\sigma_1 + \min_i d_2^i/\sigma_2)\}. \tag{3.4}$$

Defining $\bar{d}_j = \min_i d_j^i$, this generalization function becomes

$$p(y \in C|X) = \exp\{-(\bar{d}_1/\sigma_1 + \bar{d}_2/\sigma_2)\}. \tag{3.5}$$

Finally, we will replace $\bar{d}_j$ with an "effective distance" $\tilde{d}_j$ that is equal to $\bar{d}_j$ only outside the range spanned by the examples (*i.e.* when $y$ lies outside the MIN RULE bounding rectangle); inside the range of the examples, $\tilde{d}_j = 0$. This effective distance incorporates the prior knowledge that the region of healthy levels corresponds to a continuous interval along each dimension. The generalization algorithm that results from incorporating this prior knowledge along with separability into MAX SIM will be called MAX SIM$^*$:

$$p(y \in C|X) = \exp\{-(\tilde{d}_1/\sigma_1 + \tilde{d}_2/\sigma_2)\}. \tag{3.6}$$

For lack of any a priori preference between the two dimensions, we will assume $\sigma_1 = \sigma_2$ in all that follows.

In the machine learning tradition, MAX SIM or MAX SIM$^*$ are closely related to

the standard *nearest neighbor* technique for classification, in which each new object $y$ is assigned to the class with the most similar training exemplar. Of the various similarity-based models for concept learning in the psychological literature, MAX SIM has been considered to be the most flexible and robust alternative for modeling higher-level cognitive tasks (Goldstone, 1994; Medin & Florian, 1995; Osherson et al., 1990); other proposals (summed or average similarity – see section 1.4.2) would have qualitatively similar results.

Figure 3 (first two columns) shows how the two simple algorithms MIN RULE and MAX SIM* generalize from several different sets of healthy level exemplars. For MAX SIM* we plot the contours of equal probability of generalization at intervals of 0.1 between $p = 0.1$ and $p = 0.9$. The bold curve corresponds to $p(y \in C|X) = 0.5$, a natural boundary for generalizing the concept and a standard for comparison with MIN RULE. Intuitively, MIN RULE seems to provide a reasonable model when a large number of examples have been seen (row 3), but appears far too conservative in generalizing from only a few examples (row 1). Also, MIN RULE's conservative bounds seem more justified when the examples are more tightly clustered (row 4, horizontal direction) than when they are relatively spread out (row 4, vertical direction). MAX SIM* shows the opposite behavior, generalizing reasonably from a few evenly distributed examples (row 1), but generalizing just as liberally beyond the observations regardless of their number and distribution (rows 2, 3, & 4). More generally, the very idea of graded, similarity-based generalization seems most appropriate when only a few, evenly distributed examples have been observed (row 1), while all-or-none, rule-based generalization seems more appropriate after many examples have been observed to lie in a given region (row 3), or the examples have been observed to cluster tightly along one dimension (row 4).

Thus it appears that no single existing model will suffice for all cases of concept learning from positive evidence in the healthy levels domain. In other learning contexts, either MIN RULE or MAX SIM* might be perfectly appropriate for modeling generalization in a continuous feature space with separable dimensions. In particular, MIN RULE is a PAC learning algorithm for the rectangles task, which means that it

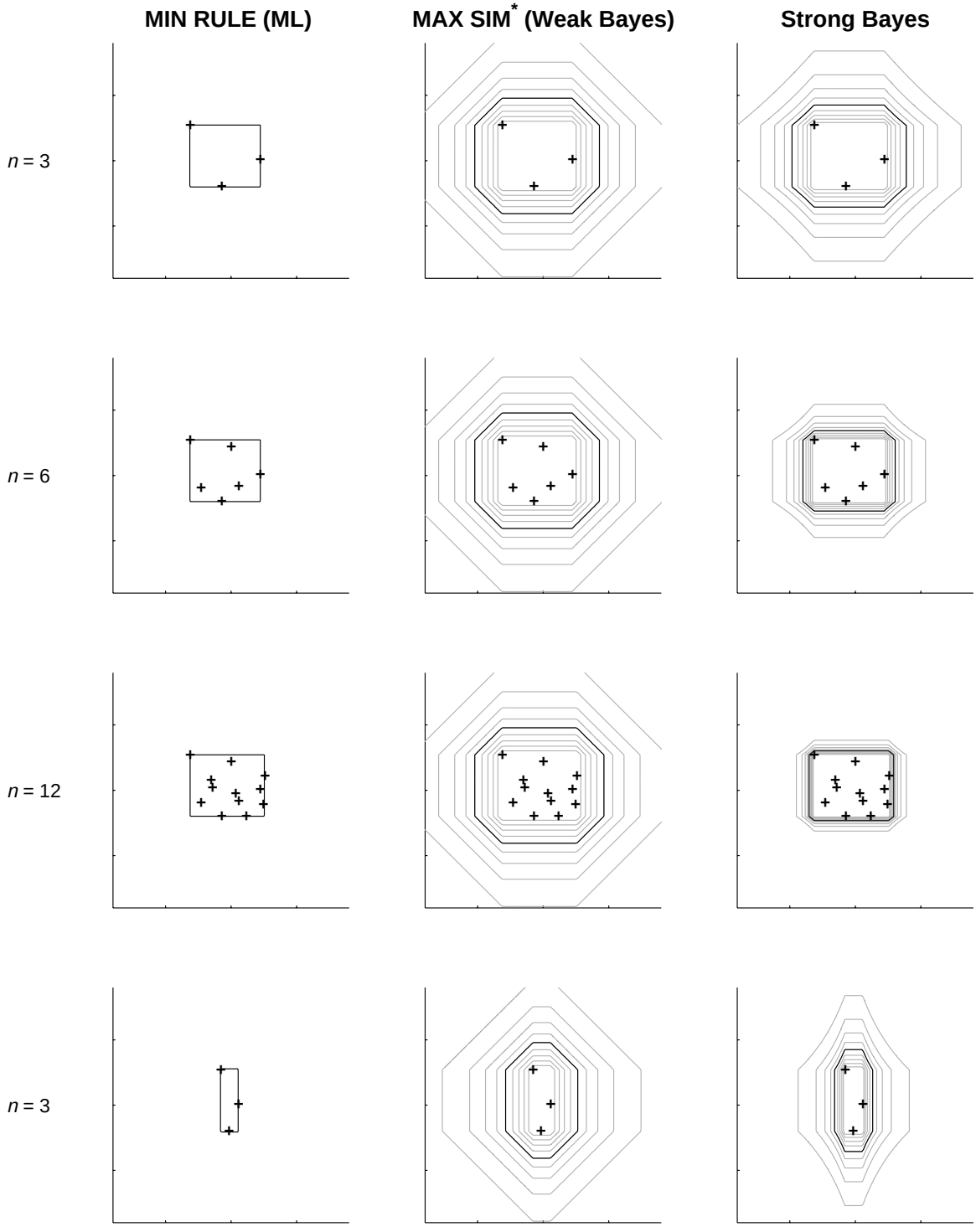



**Figure 3**

is guaranteed to get arbitrarily close to the best rectangle concept given sufficiently many (with an emphasis on *many*) examples (Kearns & Vazirani, 1994). But in this problem, over 100 examples are required to reduce the error rate below 5% with 95% confidence, according to the simple bound in chapter 1 of Kearns & Vazirani (1994). When both positive and negative examples are available, MAX SIM (or really its close cousin, nearest neighbor classification) is guaranteed asymptotically (*i.e.* as the number of examples goes to infinity) to come within twice the error rate of the optimal Bayesian classifier, and frequently does much better than expected with a "reasonable" (*i.e.* in the hundreds) number of examples (Duda & Hart, 1973). However, neither of these guarantees are of any use for the problems we are most interested in: understanding how people can learn concepts from just a few positive examples.

Figure 3 (last column) shows how the Bayesian framework for concept learning introduced in Chapter 2 combines the best of simple rule-based and similarity-based algorithms on this task. The Strong Bayes concept learner generalizes liberally and in a graded fashion where similarity-based generalization seems most appropriate (row 1), but sharpens up to more conservative, practically all-or-none behavior when rule-based generalization is called for (rows 3 and 4). The same kind of transition from similarity-based to rule-based generalization was illustrated on the number concept game in Chapter 2, Figure 8, and is a hallmark of the Bayesian approach. The Bayesian model's ability to infer how far and in what ways to generalize – whether on the number concept game, the healthy levels task, or the word learning task in the next chapter – comes from the same source: the combination of a hypothesis space of candidate rules with a probabilistic model of our observations. This allows us to weight different consistent hypotheses as more or less likely to be the true concept based on the particular examples observed. Specifically, we assume that the examples are generated by random sampling from the true concept (the *strong sampling* of Chapter 2). This leads to the *size principle*: smaller hypotheses become more likely than larger hypotheses (Figure 2b – darker rectangles are more likely), and they become exponentially more likely as the number of consistent examples increases (Figure 2c). Together with the idea of hypothesis averaging, the size principle is the

key to understanding how we can learn concepts from only a few positive examples.

### 3.2.3 The Strong Bayes model

This section gives a mathematical description of the Strong Bayes algorithm for learning rectangle concepts, and shows precisely how the MIN RULE and MAX SIM* algorithms emerge as special cases. This section features the heaviest mathematics of the thesis. As a respite in the midst of it all, I will pause halfway through and try to build up an intuitive picture of the Bayesian framework's behavior.

Recall the four components of the Bayesian framework for modeling concept learning, summarized at the end of Chapter 2:

**Summary of the Bayesian framework**

1. The notion of a constrained **hypothesis space** of possible extensions of a concept and a probability distribution over that space, representing the learner's state of knowledge about which entities a concept refers to;

2. An informative **prior distribution** over the hypothesis space reflecting the background and contextual knowledge that the learner brings to the task;

3. The **size principle** for scoring the likelihood of hypotheses under the *strong sampling* generative model, favoring smaller consistent hypotheses with exponentially greater weight as the number of observed examples increases;

4. The notion of **hypothesis averaging**, *i.e.* integrating the predictions of multiple consistent hypotheses, weighted by their posterior probabilities, to arrive at the probability of generalizing a concept to a new entity.

In Chapter 2, each of these elements was illustrated on the number concept task; here we adapt the same ingredients to the healthy levels task. The one major difference is that we are now working over *continuous* spaces of objects and hypotheses. As a result, probability distributions become probability densities and sums become integrals. But all of the machinery of probability theory carries over to continuous

spaces (Jeffreys, 1961), and up to these two changes, the mathematical derivation of the theory here will look just like Chapter 2. I should add that this derivation has much in common with Jeffreys' (1961) analysis of interval estimation and Shepard's (1987) analysis of one-trial generalization, and my presentation below owes much to the expositions of both Shepard and Jeffreys.

We have already specified our hypothesis space $\mathcal{H}$ to be the space of all axis-parallel rectangles in our two-dimensional feature space. For mathematical simplicity, we will assume for now that this feature space has infinite extent in all directions and deal with the consequences later. Then we can denote each rectangle hypothesis $h$ by a quadruple $h(l_1, l_2, s_1, s_2)$, where $l_i \in [-\infty, \infty]$ is the location of $h$'s upper-right corner and $s_i \in [0, \infty]$ is the size of $h$ along dimension $i$.

During learning, we observe $n$ positive examples $X = \{x_1, \ldots, x_n\}$ of concept $C$ and want to compute the *generalization function* $p(y \in C|X)$, i.e. the probability that some new object $y$ belongs to $C$ given the observations $X$. Our probabilistic model consists of a prior density over the hypothesis space, $p(h)$, and a likelihood function $p(X|h)$ for each hypothesis $h \in \mathcal{H}$. We first consider the likelihood, which is the more generally applicable and explanatorily central aspect of the model. Just as in Chapter 2, we will focus on the case of *strong sampling* – examples are random samples from the true extension of the concept – which leads to likelihoods inversely related to hypothesis size. In the simplest case, each example in $X$ is assumed to be independently sampled from a uniform density over the concept $C$. For $n$ examples we then have:

$$\begin{aligned} p(X|h) &= 1/|h|^n \text{ if } \forall i, x_i \in h \\ &= 0 \text{ otherwise,} \end{aligned} \tag{3.7}$$

where $|h|$ denotes the size of $h$. For rectangle $h(l_1, l_2, s_1, s_2)$, $|h|$ is simply $s_1 s_2$. [3]

[3]Here we are implicitly using the similarity metric on our feature space introduced at the beginning of this chapter. Unlike MIN RULE (see Note 1), the Bayesian model requires a metric on our space of objects (technically, only a measure), in order to be able to assign a probability density over the hypotheses (which are subsets of that space).

Note that because each hypothesis must distribute one unit mass of likelihood over its volume for each example ($\int_{x \in h} p(x|h)dh = 1$), the probability density for smaller consistent hypotheses is greater than for larger hypotheses, and exponentially greater as a function of $n$. This is the mathematical basis behind the size principle for scoring hypotheses (illustrated in Figures 2b and 2c, with darker rectangles being more likely). [4]

We now turn to the prior probability $p(h)$. The appropriate choice of $p(h)$ depends on our background knowledge. If we have no *a priori* reason to prefer any rectangle hypothesis over any other, we can choose the scale- and location-invariant *uninformative* prior (Jeffreys, 1961; Berger, 1985), [5]

$$p(h(l_1, l_2, s_1, s_2)) = 1/(s_1 s_2). \tag{3.8}$$

[4]While I stress the case of learning concepts from positive examples only, it is easy to adjust the likelihood to include the influence of negative examples as well. However, this does require that we have a generative model for the negative examples, which may be quite different from our generative model for the positives. For instance, consider the case of word learning. Adults will spontaneously use a word together with the objects it applies to, thus explicitly letting kids know that those objects are positive examples of the word. This process of generating examples can be described by the strong sampling model, as I argued in Section 2.3.4. However, adults do *not* spontaneously and explicitly give kids negative examples of a word's use; they do not (in general) go around saying things like, "Look at the cute non-doggie!", or "Would you like some non-apple juice with your applesauce?" Adults do, on the other hand, sometimes provide negative evidence about word meaning in the form of corrections to kids' misidentifications: "No, that's not a dog, that's a bear." This feedback is a form of *weak sampling* (as defined in Chapter 2) and does not give the same kind of information about the size of a concept's extension that strong sampling does, because the object labeled was not sampled from the true concept. The easiest way to incorporate negative feedback like this is just to set the likelihood to zero for any extension that includes one or more negative examples.

[5]The usual justification for the $1/s$ density as an "uninformative" prior comes from its invariance (up to a constant) under power transformations of the variable $s$. Consider a nonlinear scaling $u = s^m$ for some power $m$. Then by the chain rule, the prior becomes

$$p(u)du = p(u)ms^{m-1}ds.$$

This will be proportional to $p(s)ds$ if $p(u) = 1/u$. In other words, choosing a prior of $p(s) = 1/s$ is equivalent to choosing the same prior for any power of $s$. Because the uninformative prior, when combined with the strong sampling likelihood ($1/s^n$), leads to a posterior of the same form ($\propto 1/s^{n+1}$), the uninformative prior is said to be a *conjugate prior* for the uniform distribution of data that we assumed under strong sampling. The fact that the uninformative prior is conjugate to the likelihood we use is a convenience which makes analysis more tractable, but it does not reflect any deep correspondence between the two probability densities. It is perfectly reasonable to use the strong sampling likelihood with other priors, such as Equations 3.9 or 3.10, when they are appropriate to describe the learner's state of knowledge prior to seeing any examples.

In any realistic application, however, we will have some prior information. For example, we may know the expected size $\sigma_i$ of rectangle concepts along dimension $i$ in our domain. If that is all we know, the appropriate prior is a *maximum entropy* density (Berger, 1985), which takes the form of an exponential function:

$$p(h(l_1, l_2, s_1, s_2)) = \exp\{-(s_1/\sigma_1 + s_2/\sigma_2)\}. \tag{3.9}$$

In other situations, we may want the prior to capture additional qualitative knowledge about the possible sizes of concepts. For example, we may expect that concepts have a typical size of $\sigma_i$, *and* that concepts very much smaller ($s_i \approx 0$) or larger ($s >> \sigma_i$) than this are extremely rare. In that case, an *Erlang* density is appropriate:

$$p(h(l_1, l_2, s_1, s_2)) = s_1 s_2 \exp\{-(s_1/\sigma_1 + s_2/\sigma_2)\}. \tag{3.10}$$

All three of these densities are standard choices in the Bayesian literature for priors on scale variables like $s_i$. They are also all special cases of a *Gamma* density (with $\alpha = 0$, 1, and 2, respectively), which for one variable $s$ takes the form

$$p(s) \propto s^{\alpha-1} \exp\{-s/\sigma\}. \tag{3.11}$$

By setting the two parameters $\alpha$ and $\sigma$ appropriately, we can usually capture the basic form of our prior knowledge about the sizes of a concept's extension in a continuous feature space. These three priors are illustrated in Figure 4.

One of these priors $p(h)$ is then combined with the size-based likelihood $p(X|h)$ to compute the posterior probability $p(h|X)$ using Bayes' rule:

$$p(h|X) = \frac{p(X|h)p(h)}{p(X)} \tag{3.12}$$

$$= \frac{p(X|h)p(h)}{\int_{h'} p(X|h')p(h')}. \tag{3.13}$$

Finally, the generalization function $p(y \in C|X)$ is determined by integrating the

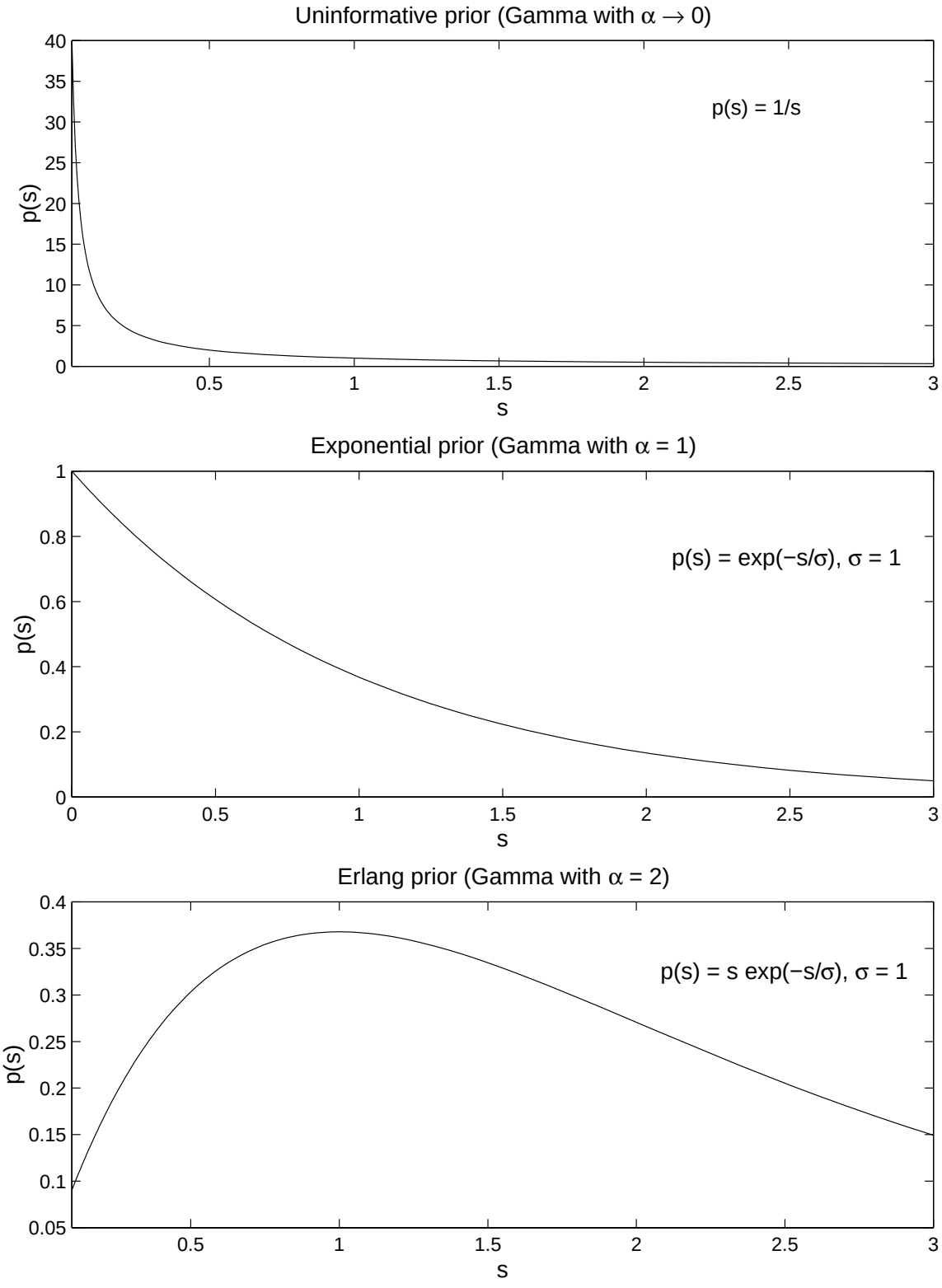

Uninformative prior (Gamma with α → 0)
p(s) = 1/s
p(s)
s
Exponential prior (Gamma with α = 1)
p(s) = exp(−s/σ), σ = 1
p(s)
s
Erlang prior (Gamma with α = 2)
p(s) = s exp(−s/σ), σ = 1
p(s)
s

**Figure 4**

predictions of all hypotheses, weighted by their posterior probabilities $p(h|X)$:

$$p(y \in C|X) = \int_{h \in \mathcal{H}} p(y \in C|h)\, p(h|X)\, dh, \tag{3.14}$$

where from Bayes' theorem $p(h|X) \propto p(X|h)p(h)$ (normalized such that $\int_{h \in \mathcal{H}} p(h|X)dh = 1$), and $p(y \in C|h) = 1$ if $y \in h$ and 0 otherwise. In order to compute the probability of generalization, we typically use the equivalent form (Equation 2.5, with sums replaced by integrals),

$$p(y \in C|X) = \frac{\int_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n dh}{\int_{h \in \mathcal{H}_X} p(h)/|h|^n dh}. \tag{3.15}$$

(Recall the notation $\mathcal{H}_X$ and $\mathcal{H}_{X,y}$ for the subsets of hypotheses in $\mathcal{H}$ that contain $X$ and $X \cup \{y\}$ respectively.) The details of the calculations are relegated to Appendix B. Here I just summarize the final answers.

Under the uninformative prior (Equation 3.8), the generalization function has a simple closed-form expression:

$$p_0(y \in C|X) = \left[\frac{1}{(1 + \tilde{d}_1/r_1)(1 + \tilde{d}_2/r_2)}\right]^{n-1}. \tag{3.16}$$

The subscript "0" denotes the fact that using the uninformative prior corresponds to a gamma prior with $\alpha = 0$. Here $r_i$ is the maximum distance between the examples in $X$ along dimension $i$. $\tilde{d}_i$ is defined to be 0 if $y$ falls inside the range of values spanned by $X$ along dimension $i$, and otherwise is the distance from $y$ to the nearest example in $X$ along dimension $i$. In other words, $r_i$ measures the size of the smallest rectangle containing $X$ and $d_i$ measures how far outside this rectangle $y$ falls, along dimension $i$. [6]

Under the exponential prior (Equation 3.9), $p(y \in C|X)$ has no simple closed-

[6]Note that Equation 3.16 is indeterminate for $n = 1$. This is reasonable under the uninformative prior; if we really have *zero* prior knowledge about the size of the concept, then seeing one example isn't going to tell us anything. Concepts of all sizes should still be equally likely, and thus the probability of generalization should equal 1/2 for all stimuli not equal to the one observed example. By taking the limit of Equation 3.16 as $r \to 0$ and $n \to 1$ (from the top) in a careful way, we can see that this is indeed the limiting case of this expression.

form expression valid for all $n$. The same is true for the Erlang prior (Equation 3.10). However, when the number of examples $n$ is greater than 2 (or when $n > 3$ for the Erlang), we can compute both an upper bound and an approximate lower bound on the generalization function using the following expression:

$$p(y \in C|X) \approx \frac{\exp\{-(\tilde{d}_1/\sigma_1 + \tilde{d}_2/\sigma_2)\}}{\left[(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)\right]^{n-\lambda}}. \tag{3.17}$$

For the exponential prior, $\lambda = 1$ provides a lower bound and $\lambda = 2$ the upper bound. For the Erlang prior, these are obtained at $\lambda = 2$ and 3, respectively. (These bounds are also derived in Appendix B.) The approximate lower bound is usually a fairly good approximation to the actual generalization function (*i.e.* within $\approx 10\%$, except for very small values of $n$ ($< 3$) and $r_i$ ($< \sigma_i/10$)), so we will use these expressions as our quick-and-dirty approximations to the probability of generalization with exponential and Erlang priors respectively:

$$p_1(y \in C|X) \approx \frac{\exp\{-(\tilde{d}_1/\sigma_1 + \tilde{d}_2/\sigma_2)\}}{\left[(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)\right]^{n-1}}, \tag{3.18}$$

and

$$p_2(y \in C|X) \approx \frac{\exp\{-(\tilde{d}_1/\sigma_1 + \tilde{d}_2/\sigma_2)\}}{\left[(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)\right]^{n-2}}. \tag{3.19}$$

Again, the subscripts "1" and "2" denote the correspondences between the exponential and Erlang priors and the gamma prior with $\alpha = 1$ and 2 respectively.

Figure 3 (right column) illustrates the Bayesian learner's contours of equal probability of generalization (at $p = 0.1$ intervals), for different values of $n$ and $r_i$. The bold curve corresponds to $p(y \in C|X) = 0.5$, the natural boundary for generalizing the concept. Integrating over all hypotheses weighted by their size-based probabilities yields a broad gradient of generalization for small $n$ (row 1) that rapidly sharpens up to the smallest consistent hypothesis as $n$ increases (rows 2-3), and that extends further along the dimension with a broader range $r_i$ of observations (row 4). This figure reflects an exponential prior with $\sigma_1 = \sigma_2 =$ half the width of the axes on the

figure; other priors produce qualitatively similar plots.

### An intuitive picture

The intuition behind why the Bayesian analysis leads to the behavior observed in Figure 3 rests on two ideas: the effect of hypothesis averaging and the size principle for scoring hypotheses. Our belief in each hypothesis $h$ as the true extension of the concept is represented by its posterior probability, $p(h|X)$. If we are quite certain that we know the extension of the concept, then only one or a few hypotheses will receive a high probability. In this case we say that the posterior probability is very *peaked* or *narrow*. If, on the other hand, we are rather uncertain about the true extension, then many hypotheses will all receive relatively equal – and relatively low – probability. In this case we say that the posterior probability is very *flat* or *broad*. Looking back at Figure 7 of Chapter 2, we can see precisely why we use terms like "flat" or "peaked". The probability distribution in the top row is literally flat, representing our uncertainty before we have seen any examples of the concept at all. After one example (row 2), the distribution develops something of a peak but it is still quite broad. After four examples (row 5), when we have a good idea of what the concept refers to, the peak has become so sharp that it falls practically to zero at everywhere but the most probable hypothesis.

Now, each hypothesis on its own predicts all-or-none, rule-based generalization of the concept. Different hypotheses instantiate different rules. When we average together the predictions of all consistent hypotheses, weighted by their probabilities, we will get out sharper or fuzzier generalization behavior depending on how the different rules are weighted by the posterior – whether it is peaked or flat, narrow or broad. When the posterior is very broad, many different hypotheses making different predictions are all averaged together with roughly equal voices. Even though each hypothesis follows a sharp rule, the effect of averaging together so many different rules is to produce a very fuzzy boundary of generalization – what looks like a gradient of similarity (Figure 3, row 1). On the other hand, when the posterior is sharply peaked, only one or a few hypotheses receive significant weight in the averaging step. These

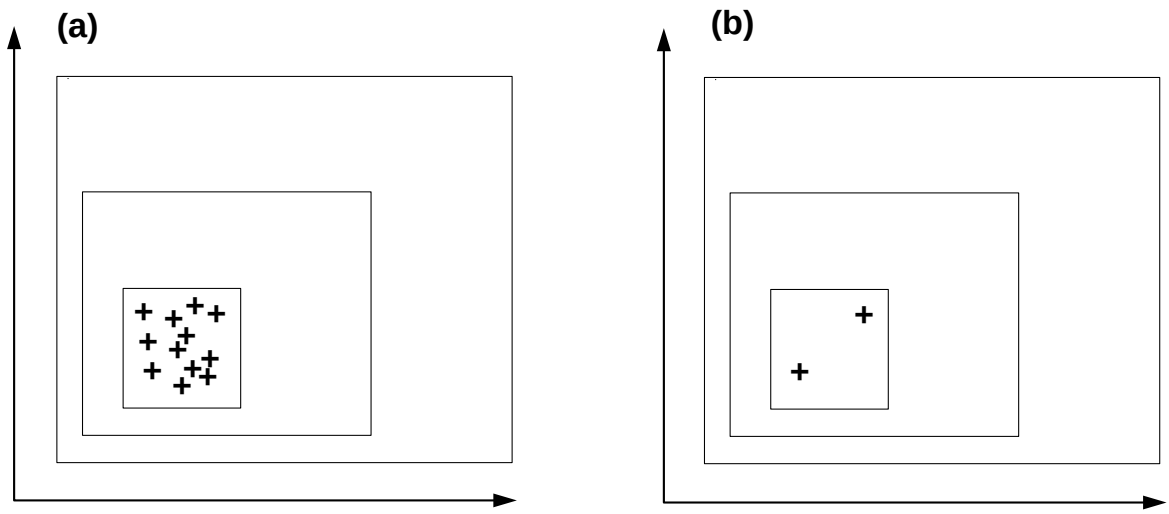



**Figure 5**

one or two loud voices drowning out all others lends the appearance of unanimity – what looks like all-or-none, rule-based generalization (Figure 3, row 3). The underlying computation is the same in both cases – integrating over a hypothesis space of possible rules weighted by probability – but the resulting behavior can look quite different. Fundamentally, it is the steepness or flatness of our posterior probability – our degree of belief in different candidate extensions – that generates the all-or-none or graded character of our generalization behaviors.

The real question, then, is what determines how steep or flat the posterior distribution will be? This is where the observed examples and the size principle enter in. The size principle, based on the assumption of randomly sampled examples, dictates that each hypothesis be assigned a probability *inversely* proportional to its size, and *exponentially* dependent on the number of examples observed. This means that small hypotheses have the advantage over larger hypotheses, and that their advantage increases rapidly as the number of examples increases. In fact, because of the size principle, the smallest hypothesis consistent with the observed examples will *always* be the most likely hypothesis. [7] However, how much more likely it is than the alternatives depends on how many examples we have seen.

[7] Although not necessarily the one with greatest posterior probability – don't forget the prior!

The size principle works exactly the same way here as it did in the number concept game in Chapter 2. Recall the intuition that it would seem very unlikely to observe the numbers 16, 8, 2, and 64 if we were sampling from all even numbers, or all numbers under 100, as opposed to only the powers of two. Observing only one even number that just happens also to be a power of two, *e.g.* 16, is still a coincidence, but far less so. Similarly, it would seem very unlikely to observe the narrow cluster of 12 healthy levels depicted in Figure 5a if they were truly drawn randomly from inside the medium-sized or the large enclosing rectangles, as opposed to just the smallest rectangle. The magnitude of the coincidence is much reduced, though, if we observe only two healthy levels in the same range (Figure 5b).

Now the key thing to realize about the healthy levels domain is that for any set of examples, there are going to be *many* hypotheses consistent with the observed examples that are all just a little bit bigger than the smallest consistent hypothesis. To put it concretely, let $h^*$ be the smallest rectangle in feature space containing the observed examples. Then there are many rectangles containing $h^*$ which are all just a little bit bigger than it. This fact has a crucial impact on the posterior probability, and thus on generalization behavior. Under the size principle, $h^*$ will always receive greater likelihood than the slightly larger hypotheses containing it, but when only one or a few examples have been observed, the likelihood preference for $h^*$ will be relatively weak. Various hypotheses of only slightly larger size will receive only slightly lower likelihood, giving rise to a posterior probability distribution that is spread roughly equally over many different hypotheses, *i.e.* a very flat distribution. Under hypothesis averaging, this will lead to a broad gradient of generalization, as in Figure 3, row 1.

However, this is only the case when a *small* number of examples have been observed. For each new example observed, the smaller hypotheses receive an additional preference over the larger ones, and these preferences *multiply* across examples. Thus, a small difference in the relative sizes of two hypotheses can become quite important after enough examples are observed, the way that the ratio of 1.1:1 becomes almost 2:1 after we multiply it by itself 7 times and almost 7:1 after we multiply it by itself 20 times. Although there will always be many consistent hypotheses that are just a

little bit bigger than the smallest consistent hypothesis, they will receive lower and lower likelihood, and thus become less and less important in guiding generalization, as more examples are observed. This narrowing of the posterior probability with each successive example is what gives rise to the sharpening of the generalization gradient that we observe in Figure 3, from row 1 to row 2, and row 2 to row 3. Convergence to all-or-none, rule-like generalization (row 3) comes when the posterior becomes so narrowly concentrated on one candidate rule that it effectively silences all other hypotheses in the averaging step.

### MIN RULE and MAX SIM* as special cases of the Bayesian framework

It is clear from Figure 3 that the Strong Bayes model combines the best of standard approaches based on rules or similarity. Given only a few examples, Strong Bayes generalization follows a broad gradient of similarity much like MAX SIM* (row 1). As the number $n$ of examples increases (rows 2-3) or their range $r$ narrows (row 4, horizontal direction), Strong Bayes converges to the most specific concept, equivalent to MIN RULE. This behavior can be understood analytically by seeing that MAX SIM* and MIN RULE correspond to different special cases of the Bayesian framework, each of which is approximately valid in different regimes of $n$ and $r$.

Recall from Chapter 2 that MIN RULE (there called MIN) is equivalent to the Strong Bayes model minus two ingredients: hypothesis averaging and an informative prior distribution. In other words, the minimal rule corresponds to the hypothesis with maximum likelihood in the Bayes framework. [8] Now, a basic property of Bayesian inference (Duda & Hart, 1973) is that it converges to maximum likelihood inference as the number of examples $n$ becomes very large (assuming every hypothesis has non-zero prior probability). Another way of saying this is that all of the posterior probability mass becomes concentrated on the single hypothesis with the greatest

[8] Hypothesis averaging, and not the particular choice of prior, is the key difference between MIN RULE and strong Bayes here. Using any of the standard priors from Equations 3.8-3.10 without adding hypothesis averaging – which corresponds to the MAP algorithm from Chapter 2 – would still lead to behavior generally equivalent to MIN RULE. However, adding hypothesis averaging in conjunction with any prior, uninformative or informative, leads to behavior that is qualitatively similar across different priors but quite different from MIN RULE (or MAX SIM*).

likelihood, with all other hypotheses receiving probability zero. [9] Mathematically, the reason Bayes converges to maximum likelihood is that the prior (Equations 3.8-3.11) is independent of $n$, while the likelihood (Equation 3.7) depends *exponentially* on $n$; thus as $n$ increases the likelihood term must eventually dominate any (finite) prior bias in its effect on the posterior. What is the implication of this fact for concept learning? MIN RULE, which neglects the prior and hypothesis averaging aspects of Strong Bayes, will nonetheless be approximately correct after we have seen sufficiently many examples of the concept.

We can see this concretely if we look at the generalization functions in Equations 3.16-3.19. All of these expressions include a term raised to the $n$th power that is equal to 1 if $\tilde{d} = 0$ (*i.e.* if $y$ falls inside the smallest rectangle containing $X$), and otherwise is less than 1. Thus, in the limit that $n \to \infty$, the probability of generalization which is the product of these $n$ terms will go to zero for all but those stimuli inside the smallest rectangle containing the observed examples (with $\tilde{d} = 0$). We can also see that each generalization function contains a factor of $\tilde{d}_i / r_i$ in the denominator. This means that when the examples are very tightly clustered along dimension $i$ (*i.e.* $r_i \approx 0$), the denominator becomes very big and the probability of generalization becomes very small *unless* $y$ falls inside the smallest rectangle containing $X$ (*i.e.* $\tilde{d}_i = 0$). Thus it is no accident that MIN RULE seems to generalize most reasonably after we have seen many examples (large $n$; Figure 3, row 3) or examples tightly clustered along one dimension (small $r_i$; Figure 3, row 4); these are precisely the cases when MIN RULE closely approximates the complete Strong Bayes model.

Unlike MIN RULE, MAX SIM was not explicitly discussed in Chapter 2's catalog of Bayesian variants. However, it turns out to be equivalent to one model that was discussed there: Weak Bayes. Recall that Weak Bayes corresponds to Strong Bayes without the strong sampling generative model, and so leads to a much weaker likelihood function than Equation 3.7. While Equation 3.7 came from assuming examples sampled randomly from the true concept, Weak Bayes assumes the examples

[9] The significance of this fact for concept learning was first noted by Watanabe (1969) and Hunt (1962).

are generated by an arbitrary process *independent* of the true concept. As a result, the size principle for scoring hypotheses does not apply; all hypotheses consistent with the examples receive a likelihood of 1 instead of the $1/|h|^n$ factor in Equation 3.7. Generalization is then determined solely by the prior and the power of the examples to eliminate inconsistent hypotheses. In the healthy levels task, using the exponential prior, the Weak Bayes generalization function is given by

$$p(y \in C|X) = \exp\{-(\tilde{d}_1/\sigma_1 + \tilde{d}_2/\sigma_2)\}. \tag{3.20}$$

Notice that this is exactly the same as the MAX SIM* algorithm (Equation 3.6). Establishing this connection between MAX SIM* and Weak Bayes was the main motivation for the simplifying assumptions we made in moving from MAX SIM to MAX SIM* earlier in this chapter. Although there is no such clean correspondence between MAX SIM itself and a Bayesian algorithm, the quantitative predictions of MAX SIM and MAX SIM* are very similar for both of the experiments presented below. [10]

This Bayesian derivation of similarity-based generalization tells us several useful things. First, it illuminates the rational basis behind the MAX SIM* algorithm; it is no longer just an equation pulled out of a hat, but can be interpreted as solving a definite computational problem under a clear set of assumptions. This analysis was motivated by Shepard's (1987) groundbreaking theory of one-shot generalization. In fact, the Weak Bayes derivation of MAX SIM* is essentially equivalent to Shepard's derivation of optimal one-shot generalization gradients extended to the case of multiple examples.

Connecting MAX SIM* with Weak Bayes also suggests an explanation of why MAX SIM* appears to generalize more reasonably in certain cases than in others: these are the cases when Weak Bayes best approximates the Strong Bayes framework. Intuitively, because Weak Bayes is equivalent to Strong Bayes in all but the likelihood term, we might expect weak Bayes to be most accurate when the influence of the

[10] The main difference between MAX SIM and MAX SIM* is that the MAX SIM generalization function has some variability within the range spanned by the observed examples, while generalization according to MAX SIM* is flat within that range. (This range was not tested in the present experiments.) Under some circumstances, this sort of variability in generalization might be quite natural; see Appendix C for a discussion of some ways it might emerge from a Bayesian learning framework.

likelihood is weakest, *i.e.* for low $n$. Comparing the generalization functions (using exponential priors) of Strong Bayes (Equation 3.18) and Weak Bayes (Equation 3.20), we see that this is indeed the case. Weak Bayes has no $n$ dependence at all. The approximate version of Strong Bayes in Equation 3.18 is actually identical to Weak Bayes for $n = 1$, and close to it for small $n$ and large $r_i$ (*i.e.* when the $1/(1+d_i/r_i)^{n-1}$ term is close to 1). Looking back at Figure 3, we see that the cases when MAX SIM looks most reasonable are just as this analysis would predict: when $n$ is small (compare row 1 with row 3) and when $r_i$ is large (compare the vertical and horizontal directions of row 4). We might also expect that in situations where the weak sampling generative model is actually appropriate to describe how examples are generated, then Weak Bayes (hence MAX SIM) will provide more generally reasonable predictions of generalization.

Finally, viewing MAX SIM* as an approximation to Strong Bayes provides a rational basis for making this simple similarity-based algorithm more flexible when it needs to be. For some time, researchers trying to ground categorization in similarity have recognized the need for some kind of flexibility in the similarity computation, in terms of how strongly different properties of stimuli are weighted in the comparison process (Goldstone, 1994; Medin & Florian, 1995). With stimuli represented as points in a continuous metric feature space, as in the healthy levels task, this flexibility is typically modeled as a "stretching" or "shrinking" of distances along the different axes of feature space (Nosofsky, 1986). This stretching or shrinking would cause generalization gradients to become compressed or expanded along these directions, much as we observe in the behavior of Strong Bayes in Figure 3. Following Nosofsky's (1986) proposal, we might model these effects in MAX SIM* by allowing the $\sigma_i$ parameters in Equation 3.5 to change depending on the examples observed, rather than being fixed a priori. (N.B. Because $d_i$ is *divided* by $\sigma_i$, a smaller value of $\sigma_i$ means that distance along dimension $i$ is weighted proportionately greater.)

But that leaves open the question of how the learner should adjust the $\sigma_i$ parameters of his similarity metric, given one or more examples of a new concept. This isn't just a minor issue of setting a model's free parameters; it is the heart of the

matter if we want to explain how people generalize concepts from just a few positive examples in terms of a "flexible similarity" construct. Nosofsky (1986) gives no formal account of how to set $\sigma_i$, although he does suggest that in general, all the $\sigma_i$ should decrease as more training examples are encountered (because the stimuli become more "distinctive"), and that $\sigma_i$ should become greater for dimensions that maximally discriminate the positive instances of the concept from the negatives (as a result of "selective attention"). The category learning models of Kruschke (1992) and Aha & Goldstone (1992) are capable of adjusting these attentional parameters in the context of a classification task, but they require both positive and negative examples to do so (in order to maximize the discriminability of positive and negative training instances). Moreover, none of these models explains *why* the feature space should be stretched and shrunk in this way.

Strong Bayes, on the other hand, does explain why these distortions of the generalization gradient occur when the learner is given only a few positive examples of the concept. In all of the expressions for the probability of generalization in Strong Bayes (Equations 3.16-3.19), the effect of the distance $\tilde{d}_i$ from a new object $y$ to the observed examples is scaled by the range $r_i$ spanned by those examples, and also becomes greater as a function of $n$, the number of examples. Thus, instead of having to postulate mechanisms of "increased distinctiveness" and "selective attention", we can understand the flexibility in generalization gradients illustrated in Figure 3 as the direct consequence of rational Bayesian inference from a given set of examples.

In sum, two major existing approaches to concept learning in continuous feature spaces, MIN RULE and MAX SIM can both be thought of as special cases of the Strong Bayes framework, corresponding to maximum likelihood (MIN) and Weak Bayes, respectively. We showed that for this problem, the maximum likelihood (MIN) rectangle closely approximates Strong Bayesian generalization for large $n$ or small $r_i$. Assuming that Strong Bayes is a good model for human concept learning on this task (as we will indeed see in the next section), this analysis explains why similarity-based generalization appears more reasonable given a few broadly spaced examples (Figure 3, row 1), while rule-based generalization appears more reasonable given

many examples (row 3) or a few examples tightly clustered along one dimension (row 4, horizontal direction). Strong Bayes can also be interpreted in the language of flexible similarity-based theories of concept generalization, as a rational prescription for how the constraints on similarity should flex given a set of observed examples. But in truth, Strong Bayes is neither strictly rule-based nor strictly similarity-based in the traditional senses. It spans both of these approaches, automatically interpolating between two regimes of "similarity-like" and "rule-like" generalization and offering the best hope for a unified theory of human concept learning. The rest of this chapter will present experimental evidence that this Bayesian framework actually describes the behavior of real human learners on healthy levels-type tasks.

## 3.3 Experiment 1

This experiment was intended to test how human learners generalize concepts in continuous feature spaces, given only positive examples of the concept. Specifically, participants were given the task of guessing two-dimensional axis-parallel rectangular concepts from one or more randomly chosen positive examples, under the cover story of learning about the range of healthy blood levels of substances like insulin, cholesterol, etc., described in Section 1. To ensure a fair test of the Bayesian framework, the experimental conditions were designed to mimic as closely as possible the assumptions made in the above theoretical analysis. The experimental procedure was designed for maximal efficiency of data collection, so that each participant was able to generalize from over 200 different sets of examples in a single 1-hour experimental session. In contrast to conventional category learning experiments, in which participants learn only one or a few categories and observe only one example at a time on each of a long series of trials, here participants learned a different concept on each trial and observed all examples of the concept at once. This procedure takes the burden off of participants' memory for exemplars and isolates the core problem of inductive generalization that is our focus here.

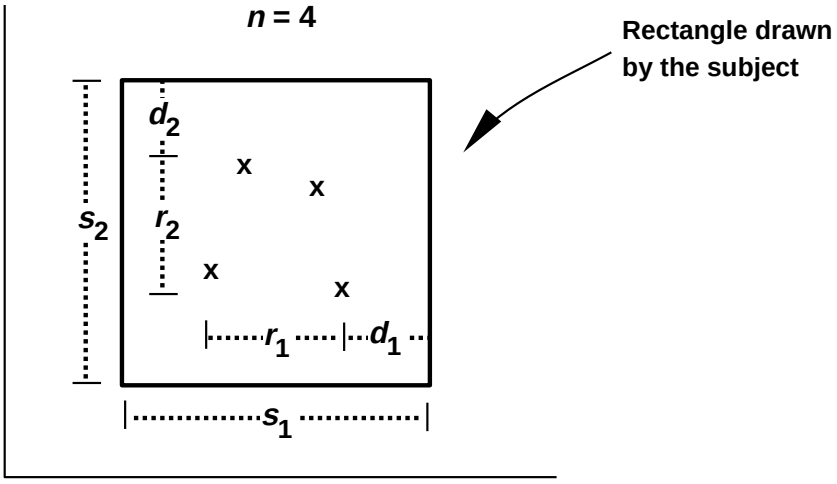
### 3.3.1 Methods

Six people participated in the study. Participants were members of the broad MIT community. All gave informed consent and were compensated for their participation. All had normal or corrected-to-normal vision.

Stimuli were presented within a 15" x 15" square window on a color computer monitor, at normal viewing distance. Participants were instructed to think of the screen as a graph on which the healthy blood levels of two different substances would be displayed as data points. Each point represented a person, and the horizontal and vertical positions of that point represented the concentrations of two substances in that person's blood. On each trial, one or more dots appeared in an otherwise blank window. Participants were told that these dots were randomly chosen points from some rectangular region of healthy levels decided upon by doctors. Their job was to guess that rectangle as nearly as possible by clicking on-screen with the mouse. For each rectangle, a participant would click three times, indicating the endpoints of the rectangle's top edge with the first two clicks and the endpoints of the right edge with the second two clicks. (The rest of the rectangle was automatically determined from this input.) Participants were instructed as follows: "Try to include *all* the points that you think could reasonably belong to the true rectangle. But don't include all the points on the screen! Try to include *only* those that you think could reasonably belong to the true rectangle." If they were unhappy with any guess after they had entered it or if they felt like they had made a mistake, participants were allowed to re-enter their guess as many times as they liked.

The stimuli were in fact randomly generated on each trial, subject to the constraints of three independent variables that were systematically varied across trials in a $(6 \times 6 \times 6)$ factorial design. The three independent variables were the horizontal range $r_1$ spanned by the dots (.25, .5, 1, 2, 4, 8 units in a 24-unit-wide window), vertical range $r_2$ spanned by the dots (same), and number $n$ of dots (2, 3, 4, 6, 10, 50). Participants thus completed 216 trials in random order. In addition, 9 trials on which only a single example appeared were randomly interspersed (but not analyzed

# Healthy levels experiment




3 independent variables:
(complete 6 x 6 x 6 factorial design, within−subjects)

- $r_1$ = {.25, .5, 1, 2, 4, 8} units out of 24 unit screen
- $r_2$ = {.25, .5, 1, 2, 4, 8} units out of 24 unit screen
- $n$ = {2, 3, 4, 6, 10, 50}

2 dependent variables:

- $s_1$
- $s_2$

Variables for analysis:

- $d_1 = (s_1 - r_1)/2$
- $d_2 = (s_2 - r_2)/2$

Figure 6

here). The horizontal and vertical exents ($s_1$ and $s_2$, respectively) of the rectangles entered by participants were recorded as the primary dependent variables. Figure 6 illustrates the relation between the three independent variables $r_1$, $r_2$, and $n$, and the two dependent variables $s_1$ and $s_2$.

To ensure that subjects understood the task, they first completed 24 practice trials in which they were shown, after entering their guess, the "true" rectangle that the dots were drawn from. Because dots were drawn randomly, the "true" rectangles that subjects saw during practice were quite variable and were rarely the "correct" response according to *any* of the theories considered here. Thus it is unlikely that this short practice was responsible for the consistent trends in subjects' behavior that we observed. I will give a quantitative version of this argument in the discussion that follows.

### 3.3.2 Results

Across trials, no consistent difference was observed between generalization in the horizontal and vertical directions. Thus, for purposes of analysis, we collapsed the data across these two directions. We also transformed from the dependent variables $s_i$ to $d_i = (s_i - r_i)/2$, the average extent of subjects' rectangles beyond range spanned by the observed examples, because this is the variable of interest in the various theories under consideration (Equations 3.16-3.20). Figure 6 illustrates the relation between $d_i$ and $s_i$. [11]

The data from six participants are shown in Figure 7, averaged across participants and across the two directions (horizontal and vertical). The extent $d$ of subjects' rectangles beyond $r$, the range spanned by the observed examples, is plotted as a function of $r$ and $n$, the number of examples. Several patterns of generalization are apparent. First, $d$ increases monotonically with $r$ and decreases with $n$. We will refer to these two phenomena respectively as the *sample variability* and *sample size* effects on generalization. Second, the rate of increase of $d$ as a function of $r$ is much slower for

[11] Strictly speaking, $d_i$ as analyzed was the average of the top and bottom (or left and right) extents of participants' rectangles, although only the top and right extents are shown in Figure 6.

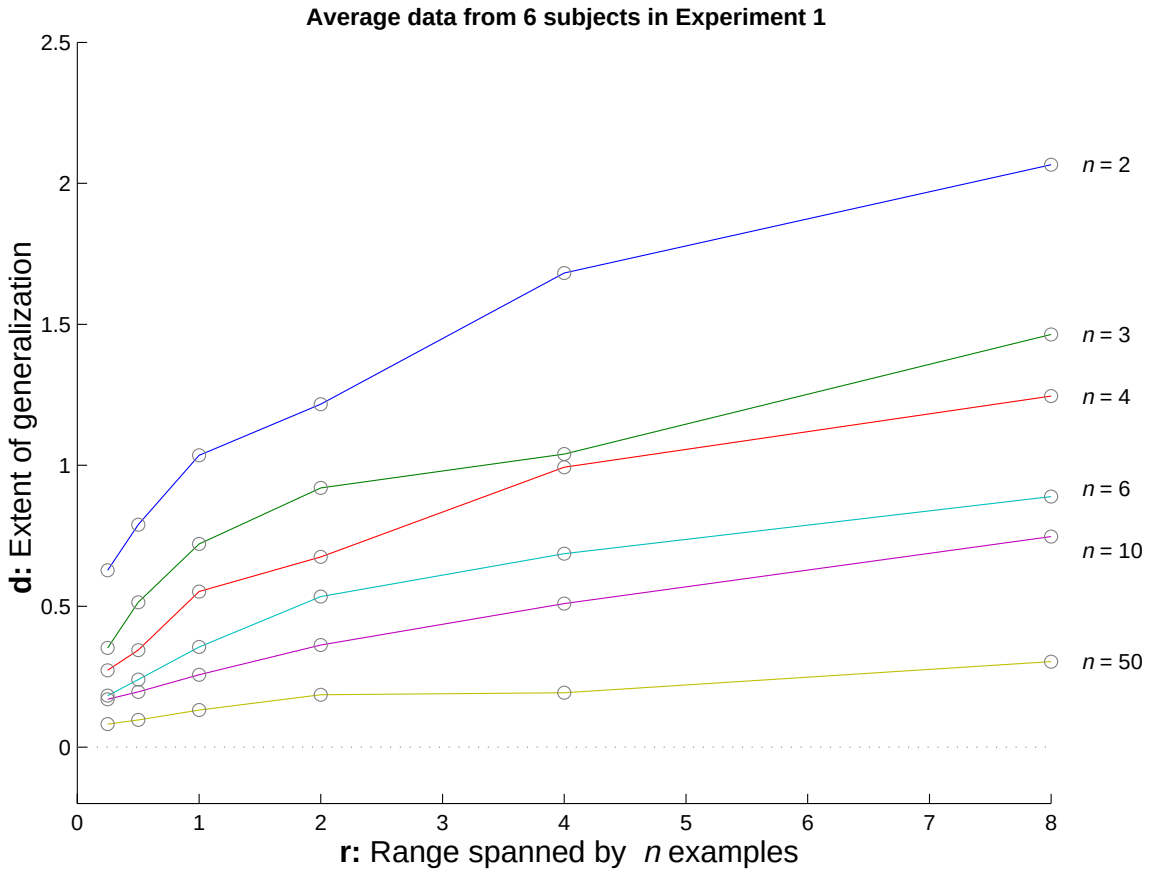



**Figure 7**

larger values of $n$. We will refer to this phenomenon as the *sample variability-sample size interaction*.

### 3.3.3 Model fits

Figure 8 confirms that neither MIN RULE (maximum likelihood) nor MAX SIM* (Weak Bayes) can explain these effects. MIN RULE always predicts zero generalization beyond the examples – a horizontal line at $d = 0$ – for all values of $r$ and $n$. The predictions of MAX SIM* (Weak Bayes) are also independent of $r$ and $n$: $d = \sigma \log 2$, assuming subjects give the tightest rectangle enclosing all points $y$ with $p(y \in C|X) > 0.5$. Any one of these horizontal lines may be reasonable for some combination of values of $r$, $n$, $d$, and $\sigma$, but clearly neither of the simple rule- or similarity-based approaches describes participants' behavior over the whole range of the experiment.

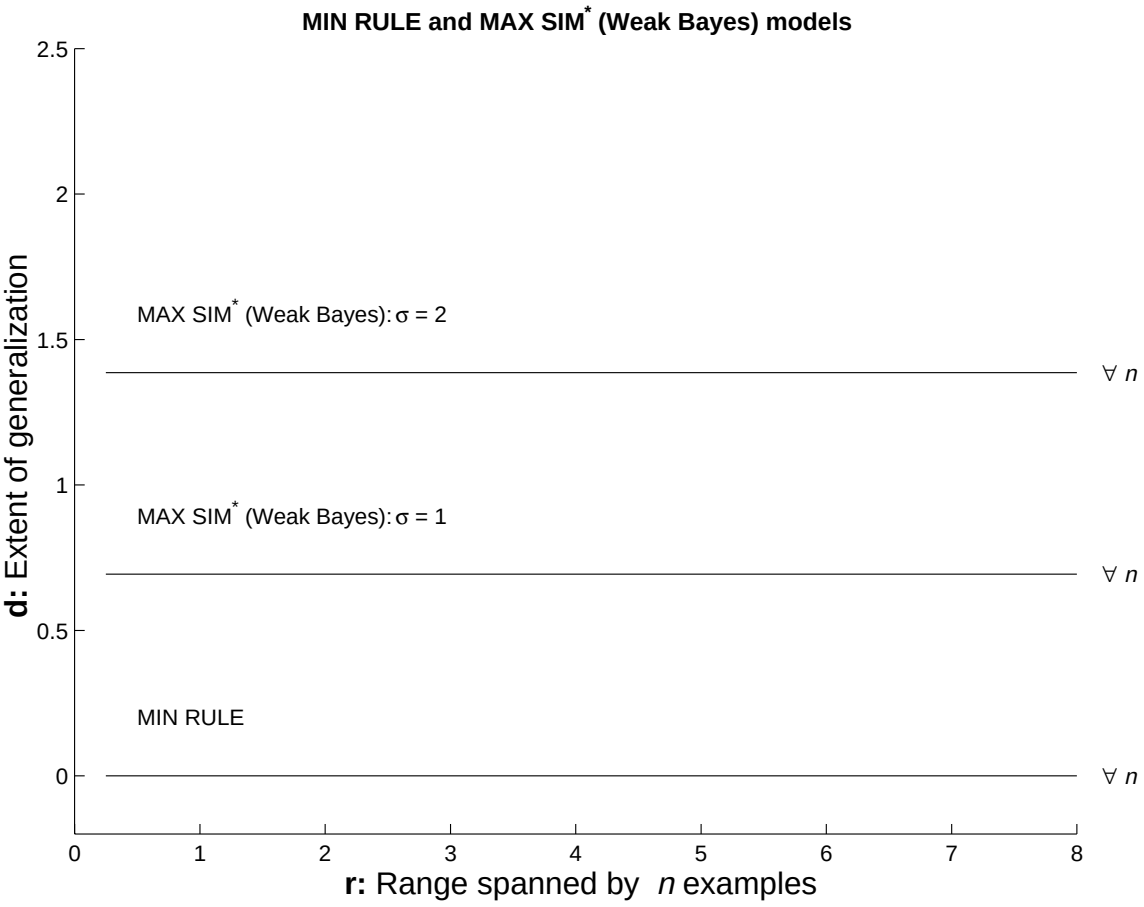



**Figure 8**

Under the same assumption, Figures 9 and 10 show the Strong Bayes model's predicted bounds on generalization using uninformative and exponential priors, respectively. Both versions of the model capture the qualitative dependence of $d$ on $r$ and $n$, including both sample variability and sample size effects, as well as a sample variability-sample size interaction. This confirms the importance of the size principle in guiding generalization independent of the choice of prior. Looking back to Figure 3, we can see these effects illustrated graphically in the predictions of the Strong Bayes model.

However, the uninformative prior misses the nonlinear dependence on $r$ for small $n$ (Figure 9), because it assumes an ideal scale invariance that clearly does not hold in this experiment (due to the fixed size of the computer window in which the rectangles appeared). In contrast, the exponential prior naturally embodies prior knowledge

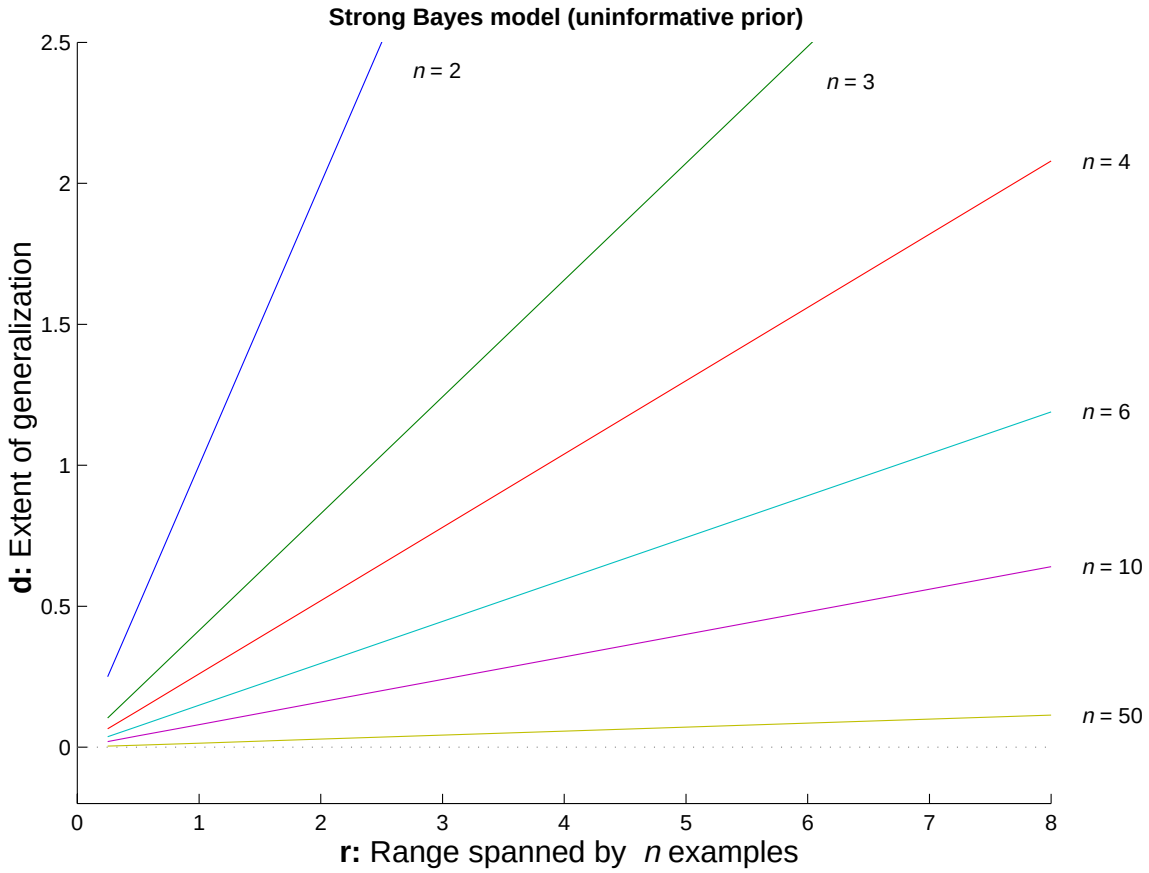



**Figure 9**

about typical scale in its one free parameter $\sigma$. A reasonable value of $\sigma = 5$ units (out of the 24-unit-wide window) yields an excellent fit to the average generalization behavior on this task (Figure 10).

### 3.3.4 Discussion

The predictions of Strong Bayes match the observed data to an impressive degree, even taking into account the one free paratemer $\sigma$. However, there are at least two alternative explanations for people's behavior, neither of which has anything to do with Bayesian models of concept learning. Let me consider each of these in turn.

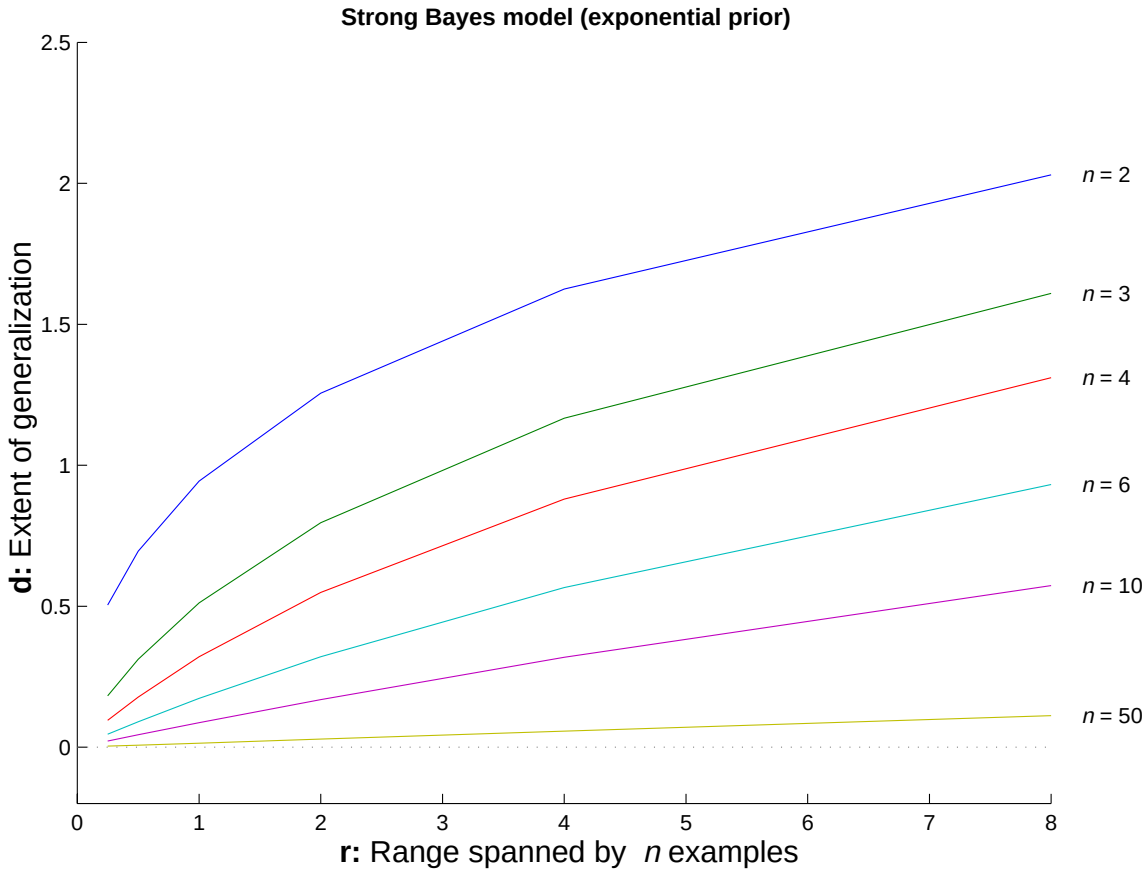



**Figure 10**

### Training artifact?

One possibility is that these results reflect a training artifact, due to the initial practice phase of the experiment. Recall that participants began the experiment with 24 practice trials in which the stimuli were generated at random from an arbitrary rectangle and this "true" concept was then shown to participants after they entered their guesses. The reason for giving participants these feedback trials was to ensure that they understood the task instructions, including what it meant to say that the observed examples were "randomly chosen" from all points inside the true rectangle (which has an ambiguous meaning in English). Could exposure to this feedback somehow have trained participants who had no such predisposition to respond in accordance with the predictions of the Strong Bayes model? That is, could participants have learned how the expected sizes $s$ of the true rectangles seen only during prac-

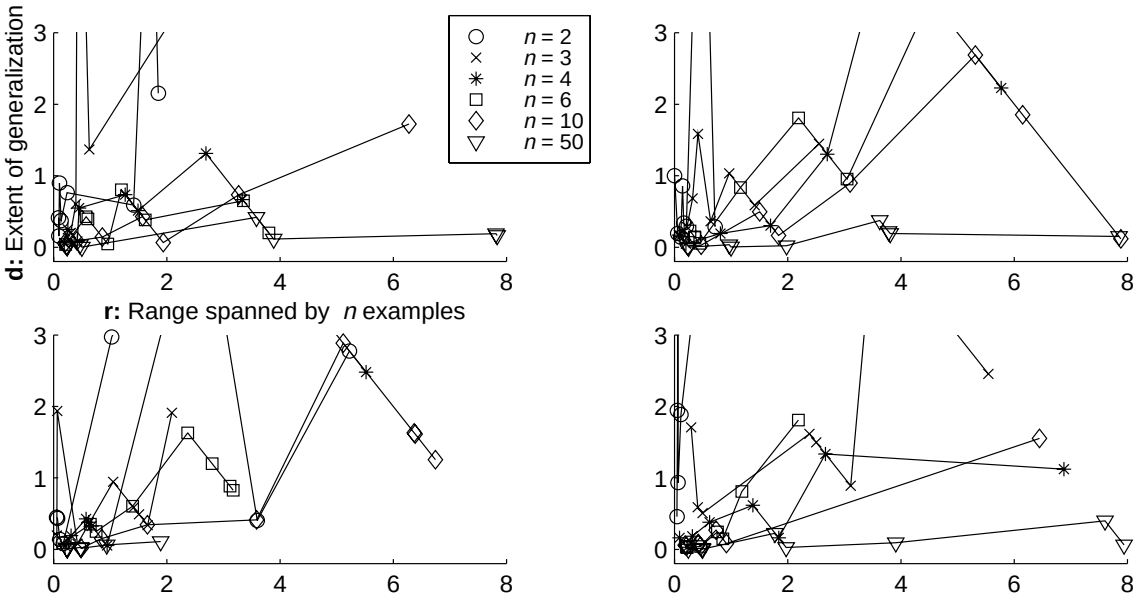



**Figure 11**

tice were related to the range $r$ and number $n$ of the examples seen on those trials, without already having an intuitive grasp of this relation?

There are several reasons not to think so. First, the practice phase consisted of only 24 trials, which is hardly sufficient to learn an arbitrary (smooth) real-valued function of two variables – *i.e.* how $s$ depends on $r$ and $n$ – unless the learner already has strong constraints on the form this function must take (Geman et al., 1992). Second, 24 trials would be a short time to learn this relationship even from perfect data, but the practice trials gave participants only very noisy information about how the expected rectangle size $s$ depends on the range $r$ and number $n$ of examples. Remember that concept learning is an underconstrained problem; thus there is *no* deterministic relationship between $r$, $n$, and the "true" value of $s$, for any one set of examples. There is a statistical relationship, but it is extremely noisy and far less consistent than the behavior of individual participants or of the participants as a whole. Figure 11 shows the raw data ($d = (s-r)/2$ as a function of $r$ and $n$) that four different participants saw in the form of feedback during their practice trials. Figure 12 shows the best-fitting linear fits to these data as a function of $r$, for each value of $n$. None of the several theoretically significant trends found in the average behavior

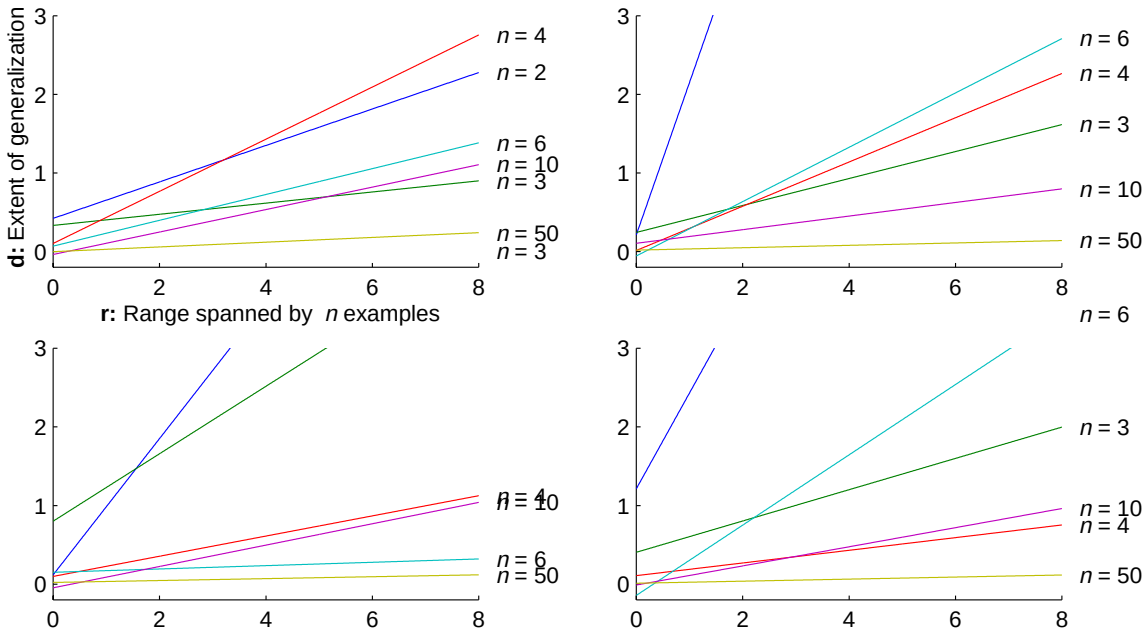



**Figure 12**

of human learners on the main part of the experiment (Figure 7) appear here at all clearly. Figure 13 shows the individual behavior of each of the six participants on the main part of the experiment. Five out of the six clearly show the same basic trends as the aggregate data. It is very hard to believe that people could have been trained to produce such consistent results (Figure 13) from such inconsistent feedback as they received during practice (Figure 11).

Certainly the practice phase had an effect on participants' behavior. In pilot experiments without a practice phase, people frequently reported that they did not understand what they were supposed to do and their behavior appeared much more erratic. But rather than "training" participants to generalize in a certain way, the practice phase seems to have "triggered" or "cued" them to use a mode of concept learning which they already had available to them. It seems reasonable that with such an artificial task as this, people would only be able to generalize meaningfully if they could activate learning algorithms normally designed for operating in much richer and more realistic environments (Gigerenzer & Hoffrage, 1995; Brase, Cosmides & Tooby, 1998; Cosmides & Tooby, 1992). Because we want to study those natural inference procedures under controlled conditions, we use an artificial task, but precede it with

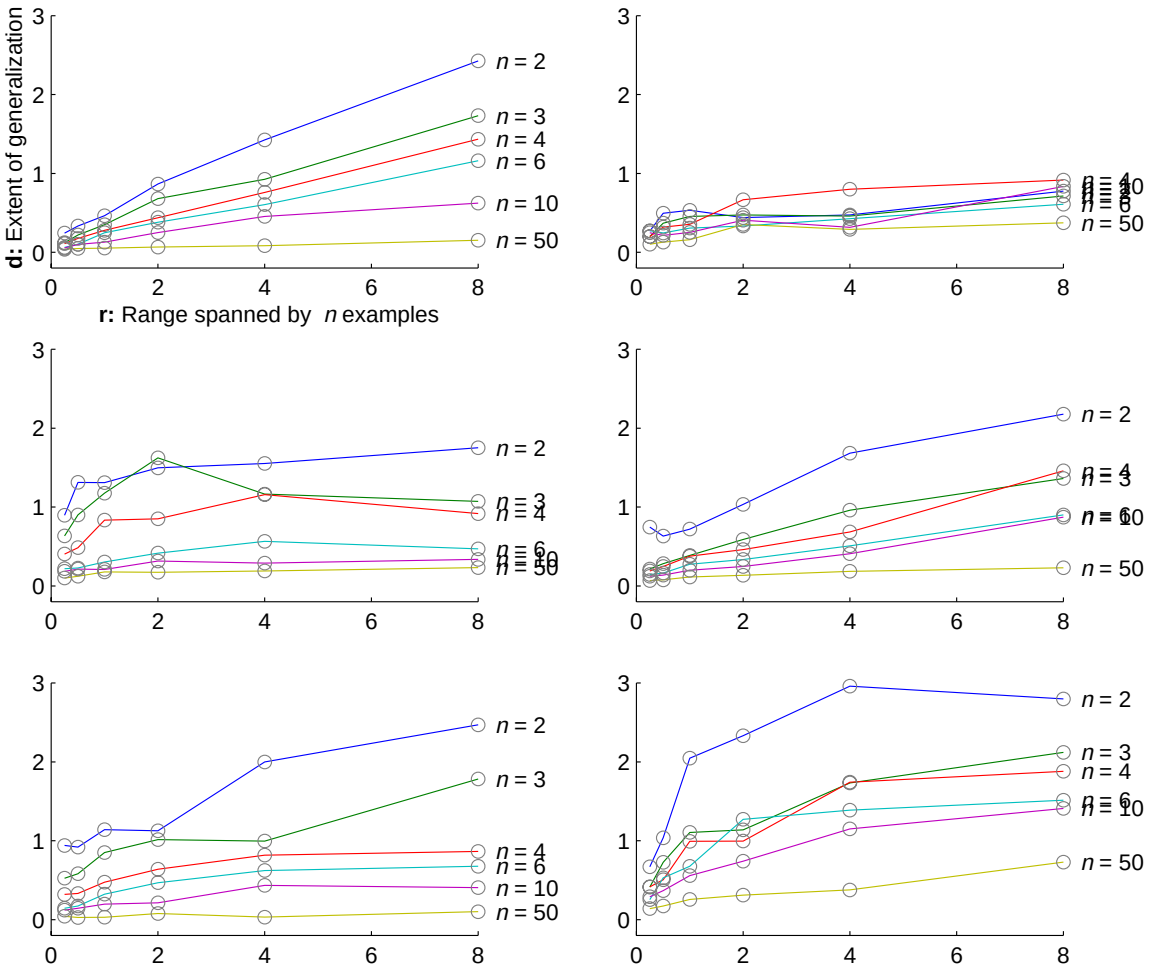



**Figure 13**

a practice phase designed to engage those procedures in a systematic way.

### Stimulus artifact

A second alternative explanation is that participants' behavior was the result of a stimulus artifact, somehow due to the graphical method in which the examples were displayed. Examples were represented visually as points in the two-dimensional physical space of the computer monitor's screen, and participants indicated their best guess at the concept by drawing a visible rectangle around the dots in that space. This interface was adopted so that participants could process all the examples of a concept quickly and at once, could easily apprehend any metric information that they chose

to attend to, and could describe their judgments of generalization by making a quick and intuitively appealing response. In pilot testing, people also found this graphical interface much more engaging and natural than other interfaces we tried. Finally, previous studies have shown that people are much better intuitive statisticians when given perceptually vivid representations of data (Cosmides & Tooby, 1996). In theoretical terms, those studies posed a much simpler task for participants than ours does; their task had an uncontroversial normative solution which required a single application of Bayes' theorem, while ours asks participants to make inferences that no existing machine learning algorithm is capable of and that require (in principle) a full Bayesian decision process over an infinite hypothesis space. Hence it seemed essential to use the most vivid representations that we could, at least for our first experiment.

Despite all these reasons behind our choice of a graphical stimulus interface, there remains the danger that as a result of this choice, our experiment will be engaging something other than people's concept learning algorithms. Is it possible that the generalization behavior we observed was not primarily the product of intuitive statistical reasoning, but rather of a spatial reasoning process, or perhaps some sort of aesthetic response? To test this possibility, we undertook a follow-up experiment.

## 3.4 Experiment 2

This experiment attempts to replicate the results of Experiment 1 using a less vivid stimulus representation, while keeping as many other details of the design the same. The major differences were as follows. Instead of representing the examples of healthy levels by points drawn on screen, we now used a straightforward numerical representation. Each example of a healthy level was now given as an integer between 1000 and 2000 on an arbitrary scale. Participants entered their estimates for the minimum and maximum healthy levels of each substance by adjusting two sliders on the computer screen, which could vary between 1000 and 2000 in steps of 1. For instance, on one trial a participant might be given two examples of healthy levels, 1410 and 1630, and

she might estimate the minimum and maximum healthy levels to be at 1350 and 1700, respectively. A perfect replication of Experiment 1 would show participants *two* sets of numbers on each trial, corresponding to the horizontal and vertical positions of the dots in Experiment 1. However, this made for a very crowded and confusing on-screen display when large numbers of examples were presented. Instead of changing the values of the independent variable $n$, we chose to show only one set of numbers per trial. From the point of view of the Bayesian framework, of course, this change from two dimensions to one is trivial – particularly since we conducted all of our analysis in Experiment 1 on data from both dimensions collapsed as if they were one.

### 3.4.1 Methods

Six people, none of whom were subjects in Experiment 1, participated in the study. Participants were members of the broad MIT community. All gave informed consent and were compensated for their participation. All had normal or corrected-to-normal vision.

Stimuli were presented within a 15" x 15" square window on a color computer monitor, at normal viewing distance. Participants were instructed as follows. On each trial, one or more numbers would appear on screen representing the healthy blood levels of different substances. Doctors had determined a minimum and maximum healthy level for each substance, and these numbers represented randomly chosen levels from within that healthy range. Their job was to guess that range as nearly as possible by setting two sliders on screen, for the minimum and maximum healthy levels respectively. Participants were told: "Try to include *all* the points that you think could reasonably belong to the true healthy range. But don't include all the levels between 1000 and 2000! Try to include *only* those that you think could reasonably belong to the true healthy range." If they were unhappy with any guess after they had entered it or if they felt like they had made a mistake, participants were allowed to re-enter their guess as many times as they liked.

The stimuli were in fact randomly generated on each trial, subject to the constraints of two independent variables that were systematically varied across trials in a

$(6 \times 6)$ factorial design. These two independent variables were the range $r$ spanned by the healthy levels (10, 21, 42, 83, 167, 333 on a scale from 1000 to 2000) and the number $n$ of examples (2, 3, 4, 6, 10, 50). The ranges were chosen to correspond exactly to those used in Experiment 1, as proportions of the total available range of stimuli. Each block of trials thus consisted of 36 trials, and participants completed three identical blocks for a total of 108 trials. In addition, one trial on which only a single example appeared were added at the end of each block (but not analyzed here). The difference $s$ between participants' estimates of the minimum and maximum healthy levels was recorded as the primary dependent variable.

Just as in Experiment 1, to ensure that participants understood the task, they first completed 24 practice trials with feedback on the "true" healthy range that the levels were sampled from.

### 3.4.2 Results

As in experiment 1, we transformed from the dependent variable $s$ to $d = (s - r)/2$, the average extent of participants' range esimates beyond the range spanned by the observed examples. The data from 6 subjects are shown in Figure 14, averaged across subjects. The extent $d$ of subjects' range estimates beyond $r$, the range spanned by the observed examples, is plotted as a function of $r$ and $n$, the number of examples. For purposes of comparison with the results of Experiment 1, the $d$ and $r$ axes in Figure 14 are scaled by a factor of 24/1000 to make all absolute values directly comparable with Figure 7. The multiplier 24/1000 comes from the fact that the total available range of stimuli on Experiment 1 was 24 units (the screen width) and on Experiment 2 was 1000 units (between the integers 1000 and 2000). In other words, a difference of 6 units (out of a 24-unit-wide screen) in Experiment 1 was equivalent as a percentage of the total stimulus range to a difference of 250 units (out of a 1000-unit-wide range) in Experiment 2.

As in Experiment 1, we found that $d$ increases monotonically with $r$ and decreases with $n$, and the rate of increase of $d$ as a function of $r$ is much slower for larger values of $n$. To illustrate concretely what these trends represent, con-

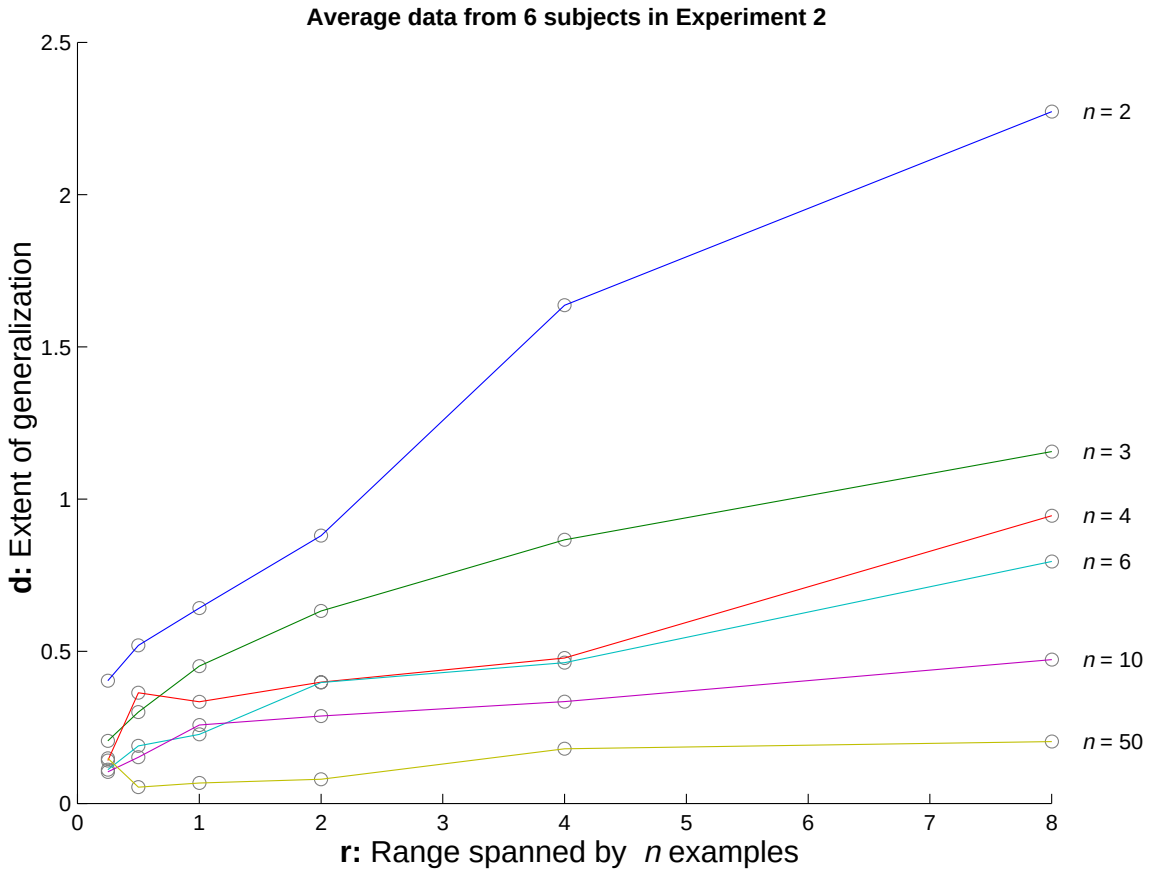



**Figure 14**

sider the following examples. Given $n$ examples of healthy levels spanning a certain range (*e.g.* $\{1400, 1700\}$), people estimated the minimum and maximum healthy levels to lie further out than they did when given $n$ examples spanning a narrower range (*e.g.* $\{1400, 1450\}$). This is the sample variability effect. Given a certain number of examples in a fixed range (*e.g.* $\{1400, 1700\}$), people generalized further than they did when given a greater number of examples within the same range (*e.g.* $\{1400, 1430, 1470, 1500, 1520, 1580, 1610, 1650, 1660, 1700\}$). This is the sample size effect. Finally, the difference in generalization between trials with different ranges and few examples (*e.g.* $\{1400, 1700\}$ versus $\{1400, 1450\}$) was much greater than the difference in generalization between trials with the same difference in ranges but greater numbers of examples (*e.g.* $\{1400, 1430, 1470, 1500, 1520, 1580, 1610, 1650, 1660, 1700\}$ versus $\{1400, 1402, 1410, 1417, 1420, 1426, 1433, 1435, 1441, 1450\}$). This is the sample variability - sample size interaction. I first discussed these effects in this form, and

presented a Bayesian model to account for them, in Tenenbaum (1997b).

### 3.4.3 Discussion

The qualitative dependence of $d$ on $r$ and $n$ in this experiment was the same as in Experiment 1, including both sample variability and sample size effects as well as an interaction between these two factors. Also, the absolute values of $d$ as a function of $r$ were remarkably similar in both experiments. There does not seem to be as strong a nonlinear dependence of $d$ on $r$ for low $n$, relative to Experiment 1. It is not clear whether this is due to the differences in stimulus format, or to the relatively small number of participants in each experiment (or to some other factor).

Overall, the major predictions of the Bayesian framework that held in Experiment 1 also held in this experiment. The nonlinear dependence of $d$ on $r$ for small $n$, which was not observed, is in any case not a central prediction of the Strong Bayes model; it arises with an exponential prior (Figure 10) but not with the uninformative prior (Figure 9). Thus, we can conclude that the excellent fit of the Strong Bayes model in Experiment 1 was not primarily due to an artifact of the graphical stimulus interface.

## 3.5 General Discussion

The last forty years have seen numerous studies of concept learning in continuous feature spaces like the healthy levels domain – both on how machines should, in principle, learn such concepts, and on how people actually do. What distinguishes the theory and experiments presented here is the focus on learning and generalizing from just a few *positive* examples of a new concept. There are many alternative models from both the machine learning and cognitive psychology literatures that I have not addressed in this chapter because they are essentially *discriminative* techniques – not capable of generalizing concepts in a principled way from positive evidence only. (I discuss the limitations of these models at length in Appendix A.) Likewise, I have not tried to address the many previous experiments on human category learning in which subjects are trained over a long series of trials to discriminate positive from

negative instances of a category, because I think they bear only limited relevance to the typical situations of natural concept learning – such as word learning – in which only one or a few labeled examples are observed before the learner must be able to generalize reasonably to novel objects. Instead, I have explored in detail a single case study – the healthy levels task – which tries to isolate the essential inductive challenge involved in learning concepts from limited positive evidence. In presenting a model of learning which is at once computationally principled and able to fit human behavior on this task precisely, I hope to have shed some light on how people are in general able to infer the correct extent of a concept from only a few positive examples.

Two components of the Bayesian model were crucial to its success on this case study: (1) a generalization function that results from averaging the predictions of all hypotheses weighted by their posterior probability; (2) the assumption that examples are randomly sampled from the concept to be learned. Averaging predictions over the whole hypothesis space explains why either broad gradients of generalization (Figure 3, row 1) or sharp, rule-based generalization (Figure 3, row 3) may emerge, depending on how peaked the posterior is. Assuming examples drawn randomly from the concept (the *strong sampling* assumption) explains why learners do not weight all consistent hypotheses equally, but instead weight more specific hypotheses higher than more general ones, by a factor that increases exponentially with the number of examples observed (the *size principle*).

As I showed in Section 2, traditional approaches to concept learning based on rules or similarity each embody only one of these two ingredients. The MIN RULE algorithm is essentially the size principle without hypothesis averaging – what I call MIN in Chapter 2. MAX SIM* is essentially hypothesis averaging without the size principle for weighting hypotheses differentially based on their likelihood – what I call Weak Bayes in Chapter 2. In the limits of very many or very tightly clustered examples, the size principle imposes such a strong preference for the smallest consistent hypothesis that taking a weighted average over all hypotheses is hardly different from just choosing the single most likely one. Under these conditions, MIN is a good approximation to the full Strong Bayes model, and pure rule-based generalization (as in Figure 3,

row 3 or row 4 (horizontal direction)) seems quite reasonable. In the other limit, when we have seen only a few broadly spaced examples, the size principle imposes only very weak preferences. Then it is almost as good to average the predictions of all consistent hypotheses weighted equally as it is to average them weighted by their size-based likelihoods. Under these conditions, Weak Bayes is a good approximation to Strong Bayes, and pure similarity-based generalization (Figure 3, row 1) seems quite reasonable.

By incorporating hypothesis averaging together with the size principle, the Strong Bayes model is able to describe generalization accurately in both of these extreme regimes, as well as to interpolate between them automatically. Strong Bayes can also be viewed in the tradition of flexible similarity-based accounts of concept learning (Nosofsky, 1986; Kruschke, 1992; Aha & Goldstone, 1992; Goldstone, 1994), as a rational explanation of how and why the similarity metric of feature space changes based on the observed examples.

In the course of testing the Bayesian framework with two experiments, I documented several robust phenomena of concept learning from more than one positive example. These include the effects of increasing *sample variability* – which acts to increase the extent of generalization – and increasing *sample size* – which acts to decrease the extent of genralization, as well as the *sample variability-sample size interaction* – which reduces the magnitude of the sample variability effect at large sample sizes. The effect of sample variability on generalization has been previously documented in the categorization literature (Fried & Holyoak, 1984; Rips, 1989). The corresponding effects of sample size or sample variability-sample size interaction have not, to my knowledge, been mentioned before. However, one of the more robust findings of intuitive statistics is an appreciation for the effect of large sample sizes on the confidence with which generalizations can be asserted (Nisbett, Krantz, Jepson & Kunda, 1983; Smith, Langston, & Nisbett, 1992). The effect of sample size that we observed is consistent with that earlier work. As participants observe more examples within a fixed rectangular region, they become increasingly confident that this region (as opposed to any larger regions which it is a subset of) is the true extension of the

concept, and thus become more conservative in generalizing outside that region.

### 3.5.1 Critiques of the Bayesian approach

Fully Bayesian learning algorithms, despite their claims to optimality and rationality, are often avoided by machine learning practitioners and criticized by psychologists for certain impracticalities. One general critique of Bayesian methods is that they may be computationally intractable for realistic problems. That is, the learner's hypothesis space may become so large that the computations necessary to maintain and average out the posterior probability distribution would require unreasonable time or space resources. This is why it is very important to be able to express the results of a Bayesian analysis in a simple closed form, as we did in Section 2. Even though we are operating over a continuous hypothesis space $\mathcal{H}$ containing an *infinite* number of distinct hypotheses, we are able to compute exact or approximate expressions for the necessary integrals over $\mathcal{H}$ without having to explicitly count up the predictions of each individual hypothesis. This tractability adds significantly to the Bayesian model's psychological plausibility as an account of human concept learning, as well as its practicality as a machine learning algorithm. Moreover, the closed-form solution is seen to depend on only a few relevant (and intuitively sensible) characteristics of the observed examples: how many examples have been seen, how wide a range they span, and how far they are from the new entity to be classified.

Another important critique of Bayesian methods is that they make strong assumptions about the possible hypotheses that can explain the data, and then have no principled way to deal with violations of those assumptions (Vapnik, 1995). That is, if the true concept does not correspond to an element of the learner's hypothesis space, then all bets – not to mention all promises of optimality – are off in Bayesian inference. In contrast, Vapnik's (1995) *Structural Risk Minimization (SRM)* approach to concept learning, and the related *Probably Approximately Correct (PAC)* paradigm (Haussler, 1988), are capable of placing meaningful bounds on the probability of error even if the true concept does not belong to the learner's hypothesis space.

Vapnik's point about the riskiness of Bayesian methods is a deep one. But far

from taking it as a point against Bayesian models of concept learning, I take it as a point in their favor. Human concept learners frequently jump to conclusions from very little evidence, on the basis of strong assumptions. Occasionally their inferences are wrong, even ridiculously wrong, as anyone who has ever observed a child learning to use words could attest to. Bayesian learning algorithms have exactly this general character, which recommends them strongly as accounts of human learning. Machine learning theorists (Valiant, 1984; Haussler, 1988; Kearns & Vazirani, 1994; Vapnik, 1995), on the other hand, have typically sought models with guaranteed upper bounds on the probability of error that will hold independently of the truth of the learner's assumptions about the concept to be learned. Unfortunately, the price of these broad guarantees is that the bounds provided by SRM, PAC, and similarly motivated learning theories are hopelessly weak from the point of view of human learning. On the two-dimensional axis-parallel rectangle learning task, for instance, well over 100 examples are required in the PAC framework to reduce the error rate below 5% with 95% confidence (Kearns & Vazirani, 1994); in contrast, most of the interesting change in human learners' generalization behavior happens between 2 and 10 examples (Figure 7). There is no way to have this cake and eat it too. If we want our learning algorithms to be able to learn concepts from just a few examples, the way that people do, then we have to be willing to accept that sometimes they will leap to incorrect generalizations, just as people do. Bayes takes the bad aspects of human learning along with the good; SRM, PAC, and other conventional theories of machine learning take neither.

That said, there are certainly assumptions in the Bayesian analysis of this chapter that drastically oversimplify the circumstances of natural concept learning. We have assumed that our observations are completely noise-free, and that no example is mislabeled as a positive instance when it is really negative. We have assumed that the extension of each concept corresponds to a single nicely shaped region in some accessible feature space, ignoring the possibility that a single concept could label several, disjoint such regions. We have assumed that we know the appropriate axes for our feature space, ignoring the possibility that stimuli might be represented merely as

points in a metric space with no specified coordinate system. We have assumed that the extension of a novel concept along one dimension of our feature space is independent of its extension along other dimensions. All of these complexities can actually be addressed in some form within an extended Bayesian treatment. In each case, we can expand the learner's hypothesis space $\mathcal{H}$ to include these additional possibilities of experience, such as noisy or mislabeled examples, non-connected extensions, arbitrary coordinate axes, nonindependent features, and so on. As in the Bayesian model described in this chapter, the size principle continues to be the key for weighing the relative likelihoods of these different possibilities, and hypothesis averaging continues to be the method for generalizing to new objects when we are unsure about exactly which possibility describes the true state of our world. Appendix C sketches some of these proposals.

### 3.5.2 Knowledge-driven versus data-driven concept learning

There is one more deep divide among researchers who study concept learning which we have yet to touch on. This is the division between *knowledge-driven* and *data-driven* views of concept learning. The knowledge-driven view emphasizes the importance of people's prior knowledge in guiding their generalizations from the very limited data of one or a few labeled examples. The data-driven view emphasizes the importance of the actual observed examples in focusing people's generalizations of previously unknown concepts. Not surprisingly, the data-driven view is most often by embraced by researchers studying learning with abstract stimuli in artificial environments (Nosofsky, 1986; Gluck & Bower, 1988; Kruschke, 1992; Aha & Goldstone, 1992), where people clearly have little or no relevant prior knowledge, while the knowledge-driven view is more often embraced by researchers studying natural concepts (Murphy & Medin, 1985), cognitive development (Fodor, 1975; Carey, 1985; Keil, 1989; Markman, 1989; Spelke, 1995), and language acquisition (Chomsky, 1986; Pinker, 1995), where the raw data appear far more impoverished than the abilities people ultimately acquire. The origins of this debate go back to Hume's (1739) analysis of induction and Kant's (1783) response, and ultimately to Plato's doctrine of recollection and

Aristotle's discussions of '$\epsilon\pi\alpha\gamma\omega\gamma\eta$ – induction.

Also, the knowledge-driven view is most often associated with rule-based proposals for learning, while the data-driven view is most often associated with similarity-based algorithms. This is probably because rule-based approaches make explicit the role of the learner's prior knowledge in the form of a constrained hypothesis space of possible extensions, while similarity-based approaches make explicit the role of the data in the form of the one or more examples to which similarity is computed. However, just as we have seen how the Bayesian framework bridges the gap between rule- and similarity-based algorithms, we can also see that it blurs the boundaries between knowledge- and data-driven views of concept learning.

The Bayesian framework takes the roles of both knowledge and data seriously. The prior probability (and of course the constrained hypothesis space) embodies our relevant prior knowledge; the likelihood term allows the data to be heard, and to "speak for themselves" when they speak with one voice. But more importantly, Bayes explains how prior knowledge and observed data *interact* in guiding generalization, which no purely knowledge-driven or purely data-driven theory can tell us. For example, suppose that in the healthy levels task, instead of judging the range of healthy blood levels of a substance like insulin, which is produced normally by the human body, we were asked to judge the range of healthy blood levels of an environmental pollutant, like lead. What effect, if any, does this change to a potentially toxic substance have on the generalizations that we make?

Suppose that doctors have determined the healthy range of blood levels of both substances, insulin and lead. Now, we are given one example of an insulin level chosen at random from the designated range of healthy insulin levels: 107 (on some arbitrary scale). What can we say from this about the healthy range for this substance, what the doctors consider minimum and maximum healthy levels? Not very much. Maybe the healthy insulin range goes from 100 to 115, or from 50 to 150, or anywhere in between. It does however seem a bit less likely to be from 100 to 1000, or 50 to 500.

Now, suppose that we are given one example of a lead concentration level chosen at random from the range of levels that doctors consider healthy: 107. (Try to ignore

the numerical coincidence.) What can we say from this about the healthy range for lead, the minimum and maximum healthy levels? In the case of a toxic substance, most people will have a much more definite feeling for its probable healthy range: the lowest healthy level is zero, and the highest healthy level is around 200 (or maybe 214?).

What happened here? Why does the seemingly minor change in context from a naturally produced bodily substance to a toxic environmental pollutant have a significant effect on the kind of generalizations we make from one example? Somehow, this change in context activates different background knowledge, which allows us to draw different conclusions from the same data. A purely data-driven account of generalization obviously has no explanation for this shift, because what we take to be the data, "107", is the same in both cases. A purely knowledge-driven account can explain why people think the minimum healthy level shifts to zero – no amount of lead in the blood is thought to be necessary for healthy living – but cannot explain why people think the maximum healthy level now appears to be around 200 or 215. Some kind of statistical, *i.e.* data-driven, reasoning is being invoked, along the lines that if the maximum healthy level were much higher, *e.g.* 1000 or 2000, then it would be rather an accident that the only example we were given was so low as 107. Moreover, this data-driven process is being invoked only under the influence of a knowledge-driven process, which says that the healthy levels of toxic substances should extend to zero while the healthy levels of natural bodily substances should not include zero.

As we see more random examples of lead levels that doctors consider healthy, the statistical data-driven process begins to dominate. Here are 7 more healthy levels, in addition to 107: 36, 68, 11, 98, 75, 49, 17. Now it seems quite plausible that the maximum level doctors consider healthy is around 110 or 115, not much less or much greater. At the other extreme, before we saw *any* examples of healthy levels, prior knowledge clearly dominates. Just ask someone: "Lead is a potentially toxic environmental pollutant. What do you think is the maximum blood concentration level of lead that doctors consider healthy? What do you think is the minimum level that doctors consider healthy?" Most people have no idea what the maximum

healthy level is, but many would say that the minimum healthy level is probably at or near zero. In the intermediate regime, when we have only seen one example of a healthy lead level, the generalizations we make reflect the joint influence of our prior knowledge and the likelihood of the data. Neither component alone can explain why we think what we do.

Capturing this interaction is the speciality of Bayesian methods. Based on our prior knowledge about the possible healthy levels of toxic substances, we restrict our hypothesis space $\mathcal{H}$ to include not all possible ranges between some arbitrary minimum and maximum values, but only those intervals containing a minimum at zero. Every other ingredient of the Strong Bayes framework – prior probability, likelihood, hypothesis averaging – is unchanged. The resulting probability of generalization (under an uninformative prior) comes out to be

$$p(y \in C|X) = \left[\frac{1}{1+\tilde{d}/r}\right]^n, \tag{3.21}$$

where $r$ is the maximum healthy level observed, and $\tilde{d}$ equals 0 if $y < r$, and $y - r$ otherwise.

While I have not yet conducted a formal experiment to test how well Equation 3.21 describes people's generalizations on this task, I believe it has fairly good prospects. Under the assumption we have made throughout this chapter that people will estimate the maximum healthy level to fall where $p(y \in C|X) = 0.5$, Equation 3.21 places the maximum healthy level after the one example of 107 at 214, after the three examples $\{107, 36, 68\}$ at 135, and after $\{107, 36, 68, 11, 98, 75, 49, 17\}$ at 117. These bounds seem intuitively reasonable to many people. In a class exercise where students were asked to solve a problem structurally equivalent to this one (but with a different cover story), people gave strikingly uniform answers. The median estimates for the bounds were 232, 140, and 115, after 1, 3, and 8 examples (all less than 107) respectively.

When asked to justify their answers, several students gave sophisticated and correct statistical arguments (this was MIT after all!), many students gave intuitive but informal statistical arguments, some gave formal but incorrect arguments, and sev-

eral could give no justification – "just gut feeling". Thus the uniformity in people's judgments of generalization was not mirrored in the kind of reasons they could give to back these judgments up. I draw two (tentative) conclusions from this demonstration. First, people's level of formal statistical knowledge has relatively little effect on their intuitive judgments of generalization, at least in this case. Second, people's intuitive generalizations of simple concepts reflect the *interaction* of their prior knowledge with the observed examples in subtle ways that are fundamentally captured by a Bayesian model. To argue that concept learning is driven primarily by knowledge *or* by data is to miss what this ability is all about.

However, if we really want to understand the role of knowledge in human concept learning, we need to turn from simplified tasks like the healthy levels scenario, where people have little or no relevant prior knowledge, to more complex and natural learning settings, where substantial prior knowledge is both available to human learners and essential for making meaningful generalizations from just a few examples of a concept. That is the mission of our next case study: to explore whether the phenomena and principles uncovered in this chapter will extend to more naturalistic domains of concept learning.

# Chapter 4

# Case study #2: Learning words

The last chapter's case study of the healthy levels scenario illustrated how the Bayesian framework can be used to explain the course of concept learning in simple continuous feature spaces and to model quantitatively the generalization behavior of real human learners. However the simplicity that makes the healthy levels problem so analytically tractable and empirically accessible is also their most obvious shortcoming. What do these results tell us about natural cases of human concept learning in the real world? The purpose of this second case study is to show that the same principles that underlie the Bayesian model of generalization in simple continuous feature spaces can be applied to understanding the more natural cases of concept learning that we are really interested in.

## 4.1 Introduction

The specific task we focus on is word learning; in particular, learning common nouns like "dog", "car", and so on, that refer to coherent categories of physical objects, from seeing examples of the words' referents. This task is, in some sense, the one we have had in mind all along, the one which we were trying to study in simplified form by looking at number concepts and ranges of healthy levels. Like those much simpler kinds of concept learning, learning words for objects requires the ability to generalize reasonably to a potentially infinite set of novel objects, and typically occurs from

only one or a few labeled examples of each word's referents (Carey, 1978; Markson & Bloom, 1997), with no explicit negative evidence available (Regier, 1995), and with the possibility of many candidate extensions – nested and/or overlapping – being consistent with any one example (Quine, 1960; Markman, 1989). These are precisely the challenges that the Bayesian framework of this thesis was designed to address.

In particular, much contemporary work on word learning is motivated by the following problem of induction (simplified from Quine, 1960). Suppose that until today, you had never heard the word "dog". All of a sudden, a dog (Rover, a black labrador) runs by and someone points at him, shouting "Look at the dog!" From this moment of experience, what can you infer about the extension of the word "dog"? The extension could, of course, be the set of dogs, but many other extensions are logically possible (*i.e.* consistent with this experience): all mammals, all animals, all four-legged animals, all labradors, all black labradors, all black things, all things with tails, all running things, this individual animal (Rover), all dogs and all horses, all dogs plus the Lone Ranger's horse, and so on. The logical possibilities are endless, but human word learners – child or adult – generally have no problem ignoring these red herrings and locking in on the true meaning of "dog" from one or a few experiences of this sort. How do they do it?

Existing theories of word learning fall more or less into the same divisions as introduced in previous chapters. Some researchers have stressed the importance of prior knowledge about possible word meanings (Clark, 1973, 1987; Macnamara, 1982; Markman, 1989; Bloom, in press; – often but not always within a rule-based learning framework – while others have stressed the importance of the concrete exemplars observed, usually within a similarity-based learning framework (Quine, 1960; Jones & Smith, 1993; Samuelson, Gasser & Smith, 1997). What follows is a highly selective review of this literature, meant only to outline the basic positions and give a few representative examples.

Researchers in the knowledge-based tradition have tried to formulate a number of *constraints* on possible word meanings that could render the learner's induction problem solvable. Implicit behind most of this work is the *hypothesis elimination* ap-

proach to induction, the classic rule-based account discussed in Chapter 1, in which the observed examples and a priori constraints work together to rule out logically possible hypotheses, hopefully leaving the learner with a single acceptable generalization after a reasonable number of examples. For learning nouns, one of the most basic constraints is the *taxonomic bias*, the assumption that new words refer to taxonomic classes in a hierarchical (tree-structured) system of natural kind categories (Markman, 1989). Given the one example of "dog" above, the taxonomic bias means that reasonable candidates for the extension of "dog" would include the subsets of labradors or dogs or animals or mammals, but not the subsets of all black things, all running things, all things with tails, all dogs plus the Lone Ranger's horse, etc. The taxonomic constraint can be viewed as a reasonable basis for a rule-based learner's hypothesis space, but given only one or a few examples of a word's referents, it still leaves room for a great deal of ambiguity, in particular about the appropriate level of generalization in the taxonomic tree.

Other, stronger constraints try to reduce this ambiguity, at the cost of dramatically oversimplifying the possible meanings of words. Under the *mutual exclusivity* constraint, the learner assumes that there is only word that applies to each object (Markman, 1989). This helps to circumvent the problem of bounding generalization without negative evidence, because it allows the inference that each positive example of one word is a negative example of every other word. That is, if we've seen Sox called "cat", then under mutual exclusivity we can assume that Sox cannot be called "dog", and hence that any extension including both Rover and Sox (e.g. mammals, animals) cannot be the extension of the word "dog" that we heard used to label Rover. A mutual exclusivity assumption may be useful for children learning their first words, but in general it is both too strong and too weak. Too strong, in that it's never true that only one word can apply to a particular object; too weak, in that it still leaves open the ambiguity of how far to generalize. (In the above case, "dog" could still refer to all dogs, or all labradors, or all black labradors, or just Rover himself, etc.)

Based on work by Brown (1958), Rosch et al. (1980), and Mervis & Crisafi (1982), Markman (1989) suggested the even stronger constraint that children (and in general,

adults, as well) assume a new word maps to not just any level in a taxonomy, but to an intermediate or *basic* level. Basic-level categories are generally intermediate nodes in a taxonomic tree, which maximize many different indices of category utility and are widely recognized throughout a culture (Rosch et al., 1980). For instance, in America, "dog" is the basic level in a taxonomic sequence that includes "Rover", "labrador", "dog", "mammal", and "animal"; although if you are a labrador-lover, "labrador" might be the basic level, and if you are Rover's master, then it might be "Rover". Whether children really have a bias towards mapping words to basic-level kinds is still a matter of controversy (Callanan, Repp, McCarthy & Latzke, 1994), but it is certainly a well-known and plausible proposal. Unlike the taxonomic and mutual exclusivity constraints, a basic-level constraint solves the induction problem in principle, because each object belongs to only one basic-level category. However, this assumption only applies to basic-level words, and indeed, is counterproductive for all the words that do *not* map to basic level categories. How do we learn all the other words we know at superordinate (higher than basic) or subordinate (lower than basic) taxonomic levels (Waxman, 1990)? Some experimenters have found that seeing more than one labeled example of a word may help childern learn superordinates (Callanan, 1989), but there have been no systematic theoretical explanations for these findings.

In the data-driven, similarity tradition, the proposals for word learning are not very different from the similarity-based approaches to concept learning in continuous feature spaces discussed in Chapter 3. Linda Smith (Jones & Smith, 1993) proposes that words are generalized first and foremost on the basis of similarity to exemplars as computed in a continuous metric feature space. To explain how we can learn words that refer selectively to different aspects of objects, she appeals to Nosofsky's (1986) proposal that the axes of this similarity space may stretch or shrink dynamically to downplay or exaggerate certain features. There is also room for pre-existing biases in this approach. Landau, Smith & Jones (1997) propose that children are initially biased to place more weight on an object's shape than on its other properties (such as color, size, etc.), presumably by exaggerating the axes of the similarity space corresponding to shape features. However, the shape bias is flexible and dynamic; it

applies to inanimate objects, but not to animate objects which are known to change shape spontaneously without altering their identity (Landau, Smith, & Jones 1997).

Finally, there are hybrid proposals that combine aspects of these different traditions. For instance, Regier (1995) showed that a weakened version of the mutual exclusivity constraint could be used within a neural network model to learn words with overlapping extensions. He demonstrated this approach on a toy problem of learning spatial terms (above, inside, left, etc.) without explicitly labeled negative examples. While the network ultimately learned the referents of a dozen or so overlapping spatial terms, it too has limitations as a model of human word learning: it requires thousands of training instances and must be trained with exemplars of multiple contrasting terms simultaneously.

In summary, the word learning literature contains many promising partial theories, but no unified formal account of how people learn words for object kinds at multiple levels of a taxonomy, given only one or a few labeled examples. In this chapter, I will argue that the Strong Bayesian model of concept learning provides the basis for such an account. The next section describes an experiment that attempts to capture the key challenges of word learning with a microworld of 24 real objects. The behavior of human learners on this task shows the effects of both prior knowledge and statistical information from the examples, and aspects of both rule-like and similarity-like patterns of generalization. Section 3 considers how these results might be explained by the models of concept learning introduced in previous chapters. I demonstrate concretely that the standard MIN RULE and MAX SIM learning algorithms cannot account for how human learners generalize words in this task, and then develop a version of the Strong Bayes model that does explain learners' generalizations quite well. Section 4 discusses the relevance of these results for theories of word learning and concept learning more generally.

## 4.2 Word learning in a microwold

A classic word learning experiment proceeds as follows (Markman, 1989; Callanan, 1989). The experimenter shows the subject a sample object and names it with a novel word, as in "See this? This is a blicket". The experimenter then shows the subject a small set of test objects and asks, "Which of these ones are also blickets?" The test items are chosen to represent different possible meanings for the new word, so that by observing which objects the subject chooses as blickets, the experimenter can infer what the subject thinks the word "blicket" means. If there is more than one trial in the experiment, the experimenter presents the subject with a new sample object, a new name, and a new set of test objects on each trial.

This classic paradigm diverges from the natural situation of word learning in a number of ways: in giving only a single labeled example, in showing only a small and carefully chosen set of test objects for generalization, and in changing the set of test objects from trial to trial. In our experiments, we [1] adapted the classic paradigm to make it more like a microcosm of real-world word learning tasks. We constructed a large set of test objects, 24 in all, which was held constant across all trials of the experiment. The test set had a hierarchical structure that mirrored in limited form the structure of natural object taxonomies in the world. Objects were distributed across three different superordinate categories (animals, vegetables, vehicles) and, within those general classes, many different basic-level and subordinate categories. Hence any one novel word was expected to match only a small fraction of the test set (two to eight objects, out of 24). On the first few trials of the experiment, learners generalized from just a single example of each new word. On subsequent trials, they were shown two additional examples of each word and again asked to generalize. With this design, we could more accurately study the natural course of word learning from one to several examples and the natural extent of learners' generalizations in a complex, hierarchical world.

Because we are interested in the extent to which similarity to examples can explain

[1] The experiments reported in this chapter were carried out in collaboration with Fei Xu.

how words are generalized, we also asked people to judge the similarity of pairs of objects used as test and example stimuli. These similarity judgments will be used in the analysis to follow.

### 4.2.1 Methods

Nine people participated in the study. Participants were members of the broad MIT community. All gave informed consent and were compensated for their participation. All were native speakers of English and had normal or corrected-to-normal vision.

Stimuli were presented within a 15" x 15" square window on a color computer monitor, at normal viewing distance. Participants were told that they were helping a puppet who speaks a different language than they do to pick out the objects he needs. [2] On each trial, learners were shown pictures of either one or three labeled example(s) of a novel (monosyllabic) word, *e.g.* "blicket" (or "dax", "pog", etc.), and were asked to pick out the other "blickets" from a test set of 24 objects, by clicking on-screen with the mouse.

**Test set.** We constructed a single test set of 24 objects meant to mirror the hierarchical structure of natural object taxonomies. Objects were distributed across three different superordinate categories (animals, vegetables, vehicles) and, within those general classes, many different basic-level and subordinate categories. For example, within the class of vegetables, there were peppers, a pumpkin, a carrot, a zucchini, and an onion; within the class of peppers, there were green peppers, a yellow pepper, and a red pepper. The test set is illustrated in Figure 1.

**Example sets.** Figure 2 shows all 12 sets of labeled examples used during the experiment. The first three sets contain one example each: a dalmatian, a green pepper, or a yellow truck, representing the three main branches of the microworld's taxonomy. One of these three trials (green pepper) is illustrated in Figure 3. Notice that the fixed test set contains objects matching the labeled example at subordinate (other

[2] While the experiment reported here was conducted with adult participants, the methodology was designed to be suitable for preschool-age participants as well. Hence the puppet (Waxman, 1990).

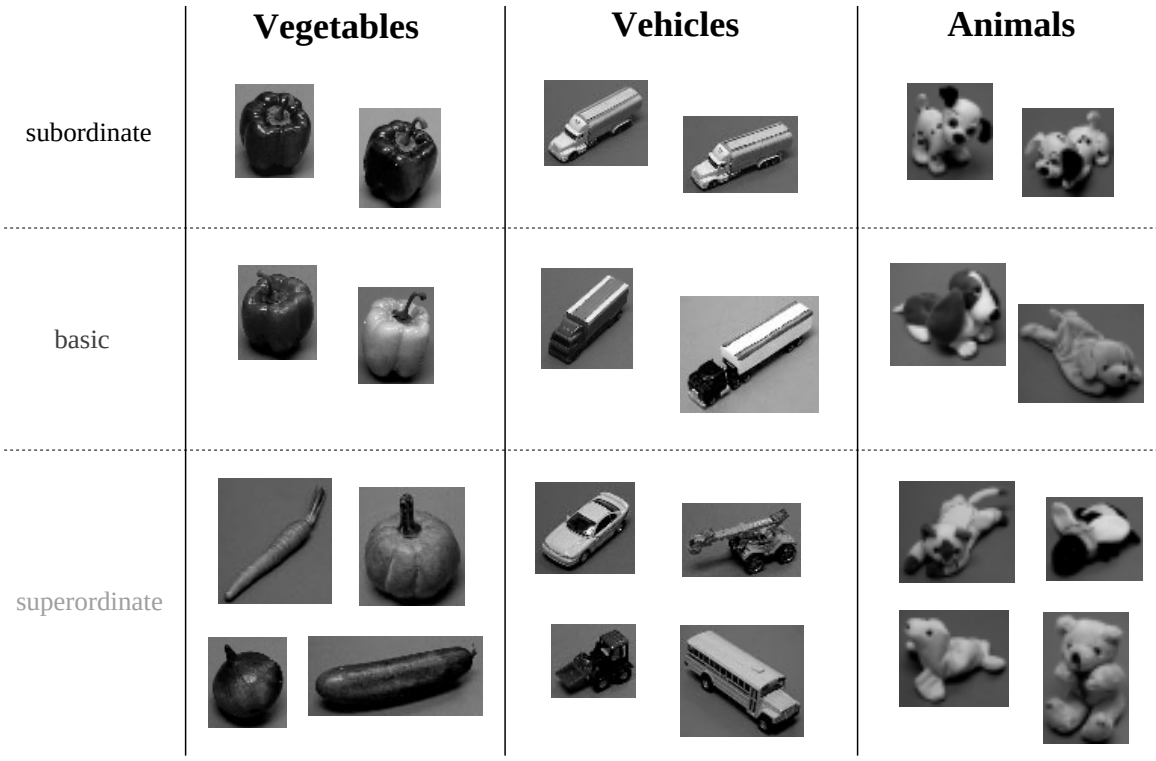



**Figure 1**

green peppers), basic (non-green peppers), and superordinate levels (non-pepper vegetables), as well as many non-matching objects (animals and vehicles). The test set is constructed so that, for any example set, there were always at least two test stimuli matching at each of these levels. In particular, the test set always contained exactly 2 subordinate matches, 4 basic-level matches, 8 superordinate matches, and 16 non-matching objects. Note that the test set is exactly the same over all trials, although the order of objects in the array is randomly permuted.

The next nine sets contain three examples each: one of the three objects from the single-example sets (the dalmatian, green pepper, or yellow truck), along with two new objects that match the first at either the subordinate, basic, or superordinate level of the taxonomy. The nine sets arise from the combination of three original objects crossed with three levels of matching specificity. Figure 4 illustrates a trial with three stimuli matching at the basic level (three peppers: green, orange, and purple). Note

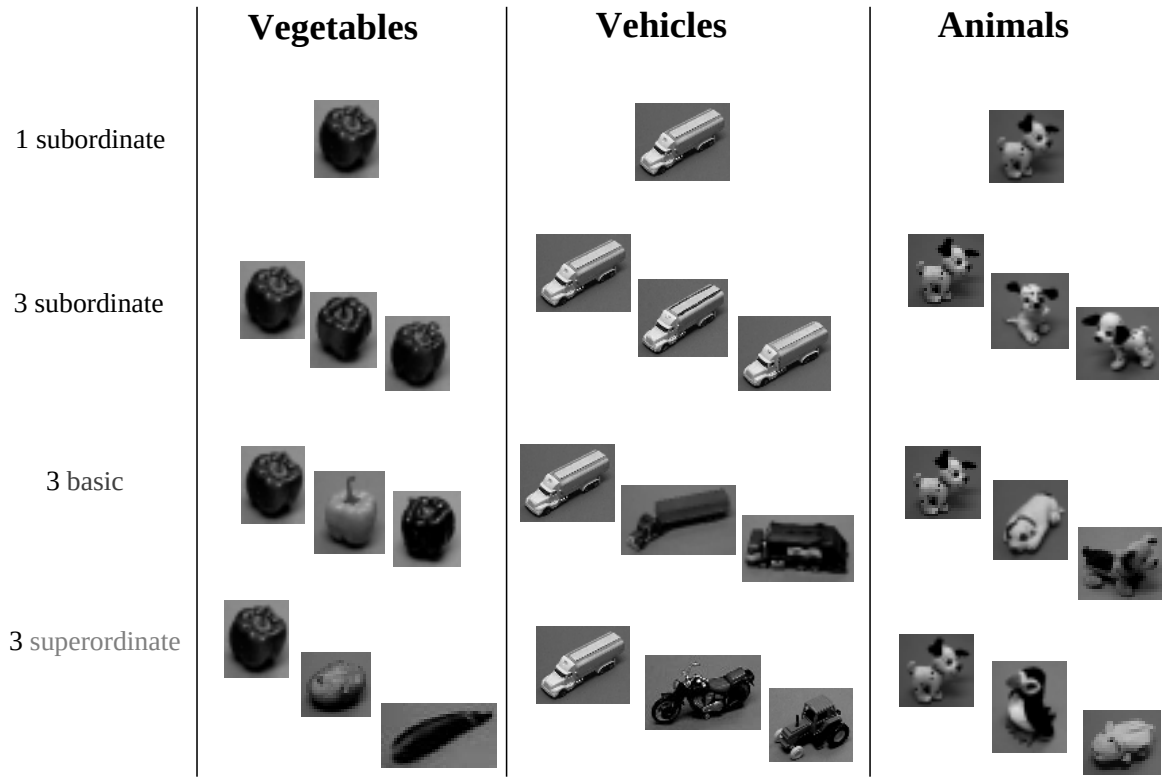



**Figure 2**

that the test set is the same as in the single-example trials, with objects matching at least one of the three examples at all different levels of specificity, along with many non-matches.

The experiment began with participants being shown all 24 test objects, one at a time, to familiarize them with the stimuli they would be working with. This familiarization was followed by the instructions and twelve experimental trials. On the first three trials, participants saw only one example of each new word, *e.g.* "Here is a blicket" (or "dax" or "pog"). On the next nine trials, they saw three examples of each word, *e.g.* "Here are three blickets." Subject to these constraints, the 12 example sets appeared in a pseudo-random order that counterbalanced the order of example content (animal, vegetable, vehicle) and example specificity (subordinate, basic, superordinate) across participants. On each trial, participants were asked to pick out


**Figure 3**

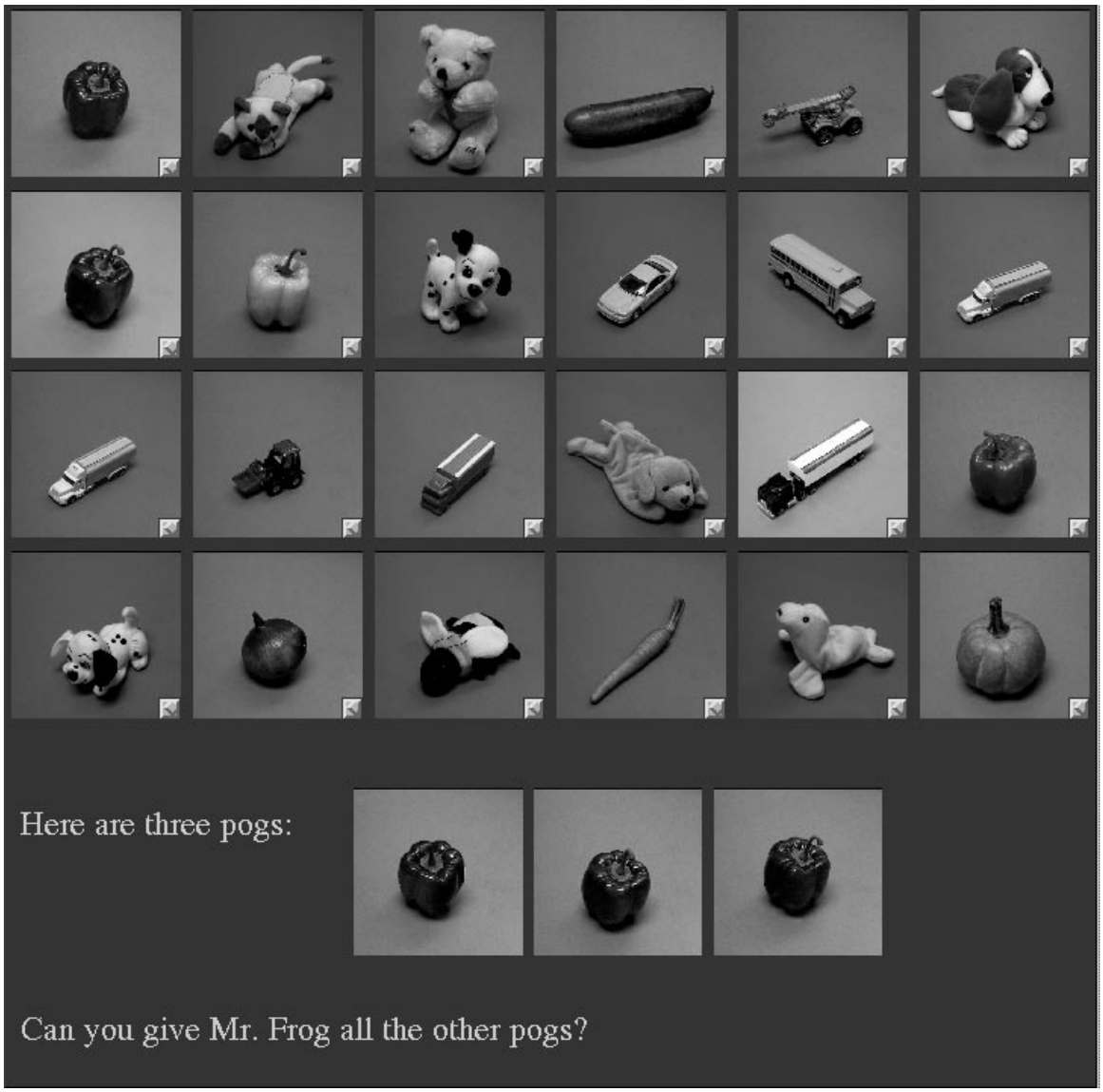



**Figure 4**

the other "blickets" from the test set by clicking on-screen with the mouse. The frequencies with which different test objects were selected during this generalization phase were the primary data. Order of responses were also collected. The same test set was used on every trial of the experiment, although the order in which objects were arrayed on-screen was randomly permuted on each trial. Following these trials, subjects completed approximately 10-15 more trials with different example sets that will not be discussed here.

Six of the participants (along with 3 participants in a pilot version of this task) also participated in a similarity judgment task following the main experiment. Participants were shown pictures of pairs of objects from the main study and asked to rate the similarity of the two objects on a scale of 1 (not similar at all) to 9 (extremely similar). They were instructed to base their ratings on the same aspects of the objects that were important to them in making their choices during the main experiment. Similarity judgments were collected for all pairs of 39 out of 45 objects used in the word learning experiment – 13 animals, 13 vegetables, and 13 vehicles – including all 24 test objects and all but six of the possible training objects (which were omitted to save time). The six omitted objects (two green peppers, two yellow trucks, two dalmatians) were each practically identical to three of the 39 included objects, and each was treated as identical to one of those 39 in constructing the model of learning reported below. Each participant rated the similarity of all pairs of 13 animals (13*12/2 = 78 judgments), all pairs of 13 vegetables (78 judgments), and all pairs of 13 vehicles (78 judgments), along with one-third of all possible cross-superordinate pairs (*e.g.* animal-vegetable, vegetable-vehicle, etc.) chosen pseudo-randomly (13*13 = 169 judgments), for a total of 403 judgments per participant. The order of trials and the order of stimuli within trials was randomized across participants. These trials were preceded by 30 practice trials (using a random sample of the same stimuli), during which participants were familiarized with the range of similarities they would encounter and were encouraged to develop a consistent way of using the 1-9 rating scale. They were also encouraged to use the entire 1-9 scale and to spread their judgments out evenly across the scale. Finally, similarity ratings for all 9 participants

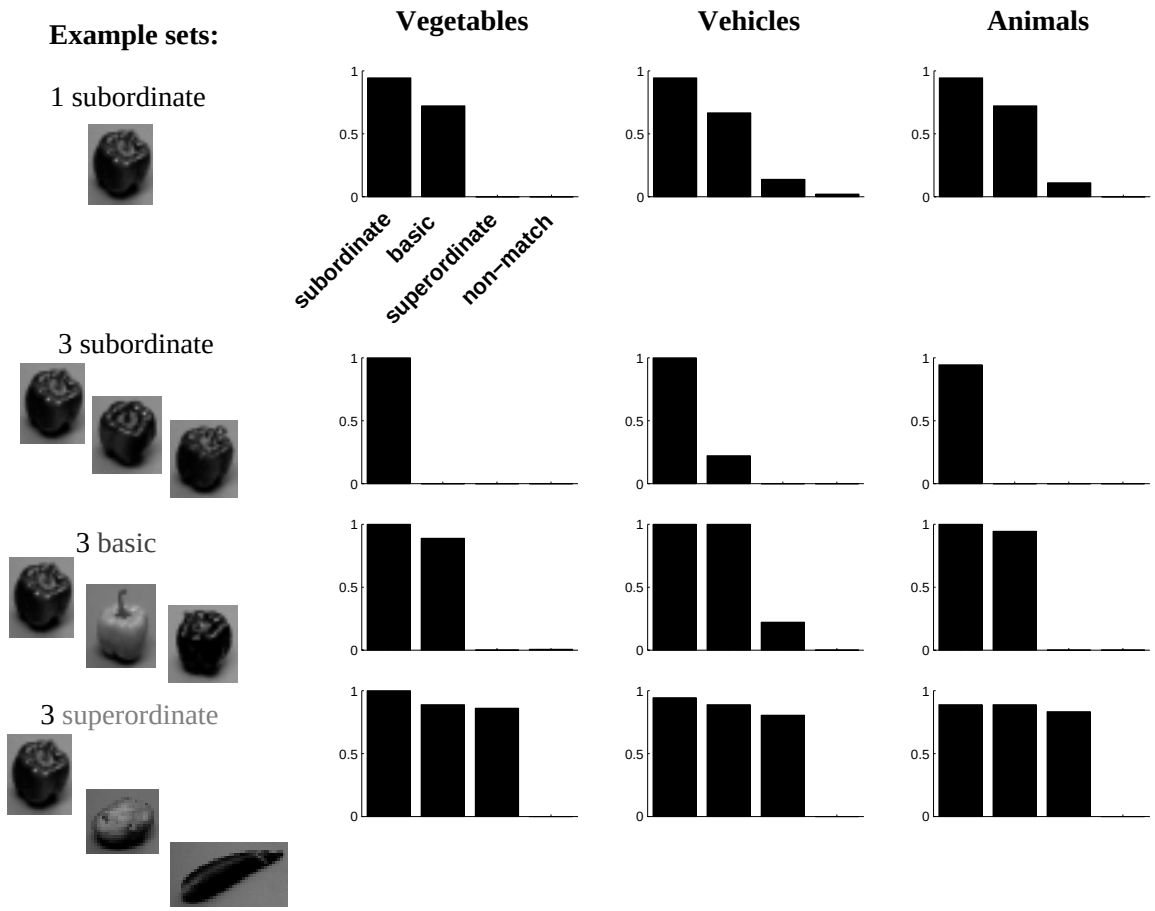



**Figure 5**

were averaged together.

### 4.2.2 Results

Figure 5 shows the generalization frequencies for all 12 example sets and all 24 test objects used in the study. Each of the 12 subplots corresponds to one set of examples; the subplots are grouped into columns according to example content (vegetables, vehicles, or animals) and grouped into rows by level of example specificity (1 example; 3 examples, subordinate level; 3 examples, basic level; or 3 examples, superordinate level). Within each subplot, there are four bars representing the frequencies of generalization from the given example set to test objects that matched the examples at one of four different levels of specificity: subordinate, basic, superordinate, or no match. Here is a concrete illustration: In the upper-left-hand-corner subplot (*i.e.*

row 1, column 1), the example set consists of a single green pepper, while the four bars represent frequencies of generalization to green peppers (2 in test set), other (non-green) peppers (2 in test set), other (non-pepper) vegetables (4 in test set), and all other objects (16 in test set).

Some broad patterns are apparent. First consider the top row of subplots, representing trials in which only a single labeled example was provided. Collapsing across all three kinds of example content, participants almost always (94% of trials) generalized to test objects matching the example at the subordinate level (*e.g.* green peppers), often but not always (70% of trails) generalized to basic-level matches (*e.g.* non-green peppers), and rarely (8% of trials) generalized to superordinate matches (*e.g.* non-pepper vegetables). Generalization to nonmatching test objects (*e.g.* animals or vehicles) was practically non-existent ($< 1\%$ of trials). Within each kind of example of content, differences between subordinate and basic matches, basic and superordinate matches, and superordinate matches and nonmatches were all significant ($p < .05$, one-tailed tests), except for superordinate vs. nonmatches with the vegetable examples. Thus, generalization from one example appears to fall off according to a gradient of exemplar similarity.

A very different pattern emerges in the bottom three rows of subplots, representing trials on which three labeled examples were provided. Instead of a gradient of generalization decreasing from more similar to less similar test objects, there appears in most cases to be a sharp transition from perfect or near-perfect generalization to practically zero generalization. The cut-off occurs at the level of the most specific category containing all three labeled examples. That is, given three examples in the same subordinate (or basic, or superodinate) class, participants generalized to all and only the subordinate (or basic, or superordinate) test matches (*e.g.* all and only the green peppers, the peppers, or the vegetables). Quantitative comparisons (collapsed across the three kinds of example content) support this claim. For three examples in the same basic category (*e.g.* three different peppers), there was no significant difference in frequency of generalization to test objects matching these at the subordinate versus basic levels. For three examples in the same superordinate category (*e.g.* three

different vegetables), there were no significant differences in frequency of generalization to test objects matching these at the subordinate, basic, or superordinate levels.

The only exceptions to this pattern of all-or-none generalization occur with three examples of vehicles at the basic and subordinate levels. Because these are not dramatic exceptions, I will set them aside for now, but I will return to them later when discussing possible models for these data.

### 4.2.3 Discussion

Our main results are consistent with the existing literature on word learning. [3] First, we found that people often but not always generalized from a single labeled example to other objects within the same basic-level category, and rarely to objects outside the same basic-level category. Second, we found that giving participants more than one labeled example had a dramatic effect on how they generalized to new objects; they tended to select all objects at the most specific taxonomic level spanned by the examples and no objects beyond that level.

The first finding is consistent with the proposal that children have a preference for mapping words onto basic-level categories (Markman, 1989; Mervis & Crisafi, 1982). On the other hand, this interpretation is complicated by the fact that our participants already knew a very familiar word in English for each of these basic-level categories, "pepper", "truck", and "dog". The tacit knowledge that objects are almost always named spontaneously at the basic level (Rosch et al., 1980) may have increased participants' propensity to map words in a new language onto basic-level categories, *i.e.* by suggesting that these new words should be translated into already familiar basic-level words in English. Because the basic-level naming bias is particularly strong for single objects rather than collections of objects (Markman, 1989), we would expect such interference to show up particularly when only one labeled example was provided. This bias in adults' tacit knowledge of how basic-level

[3] All of these connections should be interpreted with caution, because our studies employed adult participants while most word learning studies work with 3- to 5-year old children. See the general discussion of this chapter for remarks on this point.

words are used in naming could exist over and above any preference children or adults have to map new words onto basic-level categories.

The second finding is consistent with studies of how children learn superordinate words (Callanan, 1989; Golinkoff et al., 1998). These studies have found that providing two labeled examples from different basic-level categories significantly increases the frequency of generalization to other objects of the same superordinate category, relative to when only a single example is provided. Callanan's (1989) design, in particular, was similar to ours in that she had learners choose from a test set of 12 stimuli that matched the examples on multiple levels of specificity. However, in her study, the test set changed from trial to trial, each participant only saw *either* pairs of examples *or* single examples, and pairs of examples never belonged to the same basic or subordinate class. Each of these differences reduces the ambiguity in generalization that the learner faces.

Our results go beyond the existing literature in two important ways. First, we found a qualitative difference in generalization from one labeled example versus several labeled examples. While generalization from a single example decreased significantly with each decrease in similarity to the test objects (Figure 5, row 1), generalization from three examples typically followed an all-or-none pattern with a sharp threshold (Figure 5, rows 2-4). Second, we found that people used multiple labeled examples of a new word to lock in its extension at the appropriate level of specificity in a multi-level taxonomy of objects, corresponding to the most specific class containing all the examples.

## 4.3 Models of concept learning applied to learning words

As our first candidates for modeling these data, we consider the standard rule- and similarity-based approaches from last chapter, MIN RULE and MAX SIM. Although we ruled out these models as accounts of the healthy levels task in Chapter 3, it is

possible that one of them might be more suited to the more natural – and certainly more interesting – task of word learning. However, as we will see, that is not the case. After we understand the difficulties each of these models faces, we will then show how they can addressed in a unifying Bayesian framework.

### 4.3.1 Ingredients: a similarity metric and a hypothesis space of candidate rules

None of these models can be assessed, however, until we have specified some additional ingredients. Rule-based approaches require a hypothesis space of candidate rules. Similarity-based approaches require a measure of how similar any two objects are.

We obtained the similarity measure directly from the participants in the study. After completing the main experiment, they were asked to rate the similarity of many pairs of objects on a scale from 1 to 9 (see Methods section above). These ratings were averaged across subjects and scaled to the interval $[0, 1]$, to produce a normalized measure of similarity for all pairs of objects.

There are several possibilities for deriving a hypothesis space of candidate rules. The simplest approach would be to just assume that people use exactly the hypotheses that we used in designing the stimuli, corresponding to subordinate, basic, and superordinate categories for each of the three kinds of example content. Concretely, this translates into just nine hypotheses: vegetable, vehicle, animal, pepper, truck, dog, green pepper, yellow truck, and dalmatian (Figure 6). However, while this may be fine as an ideal model, it hardly exhausts all the hypotheses people could have brought to bear on this task. There are other natural candidates at every level of the taxonomy, corresponding to mammals, pets, food, legumes, four-wheeled vehicles, toys, etc., not to mention hypotheses about cats, cars, onions, etc., which are not consistent with any of the example sets but which the learner must nonetheless consider going into each trial. Also, there may be classes for which we have no simple name in English, but which are nonetheless psychologically natural candidates for the extensions of new words. The fact that different languages chop up the world in

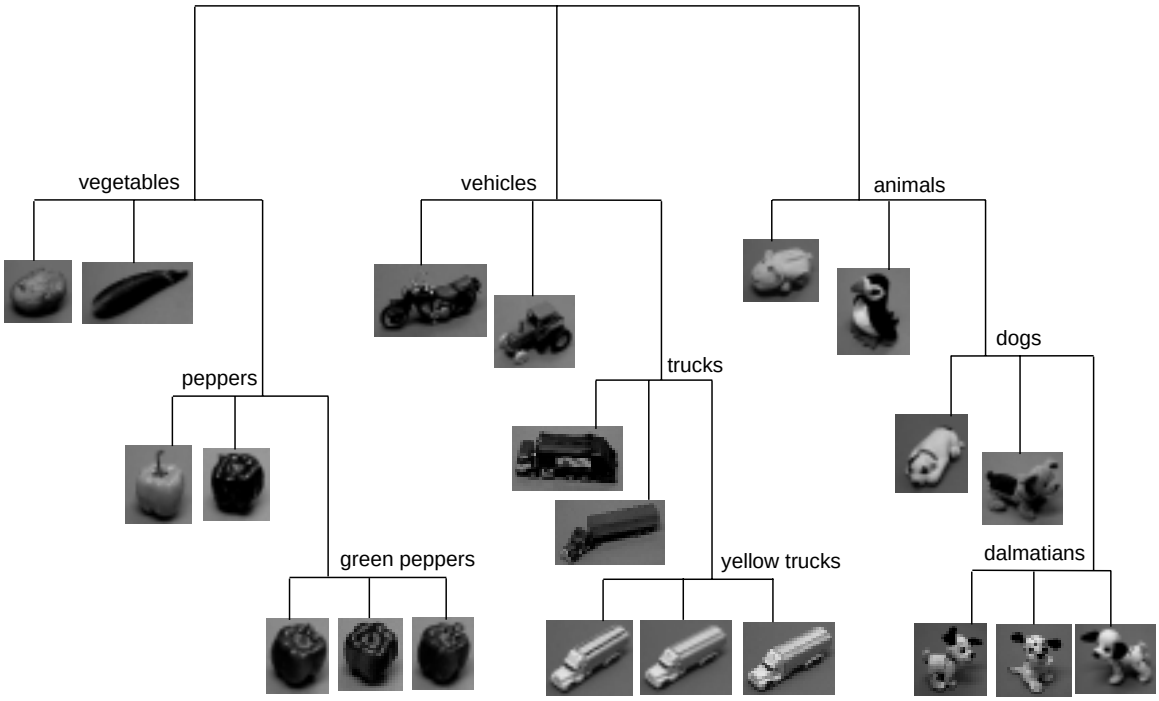



**Figure 6**

different ways, and that people are capable of learning new languages, suggest that there might be many hypotheses of exactly this sort lying around in our heads and just waiting to be triggered by new words. For all of these reasons, it would be desirable to construct a hypothesis space of candidate word extensions that was not so dependent on our superficial and linguistically biased introspections.

A more objective approach would be to ask a separate group of subjects to pick out subsets of the total set of objects that seem like natural candidates for being the extension of a word. This parallels the feature-listing methodology frequently used to constrain models of similarity based on feature overlap (Osherson et al., 1990; Sloman, 1993; Smith, 1995). However, by relying on people to specify their hypothesis space directly, this approach still confounds people's intuitions about how they learn words with what we are trying to study, how they really learn words.

We pursue a more indirect approach, looking for a hypothesis space of rules in, of all places, the similarity judgements that we collected. Recall from Chapters 1 and 2 that feature-based models of similarity can actually be thought of as *rule-*

*based* models, if we identify each feature with a rule that picks out all and only the objects having that feature (*i.e.* RED → "x is red"). Then computing similarity by counting up the number of features two objects have in common (Tversky, 1977) is equivalent to counting up the number of rules that apply to both of them. Under this view, the similarity ratings that we collected for every pair of objects $x$ and $y$ are an approximate measure of how many rules that, in our participants' minds, apply to $x$, also apply to $y$ (or vice versa).

There are a number of computational procedures for reconstructing the set of rules in peoples' minds that would give rise to a particular set of similarity judgments in this way. These include *additive clustering* (Shepard & Arabie, 1979; Tenenbaum, 1996), *hierarchical clustering* (Duda & Hart, 1973), *additive trees* (Sattath & Tversky, 1977), and *extended trees* (Corter & Tversky, 1986). Each technique embodies somewhat different assumptions about the possible structure of features/rules. Here, we adopt one of the simplest and oldest techniques, *average-link agglomerative hierarchical clustering*, or *average-link clustering* for short (Duda & Hart, 1973). Average-link clustering constructs a tree representation of the stimuli, in which each node of the tree corresponds to a cluster of stimuli that are in some sense more similar to each other than to other stimuli. Applying this technqiue to the similarity data we collected yields the tree in Figure 7.

The cluster tree has several important structural features. The nodes are depicted as horizontal lines, while the vertical lines are called "branches". Each "leaf" of the tree, where a branch terminates without a node, corresponds to one object used in the study. (For clarity, only objects that appeared as labeled examples during the main experiment are actually shown.) The height of each node represents the average pairwise similarity of the objects in the corresponding cluster, *i.e.* the objects corresponding to leaves that come under that node. (Lower height equals greater average similarity.) The length of the branch above each node measures how much *more* similar on average are the objects within it compared to objects in the next nearest cluster, *i.e.* how distinctive that cluster is. It may be more intuitive to think of distinctiveness in terms of distance, as a measure of how well separated the

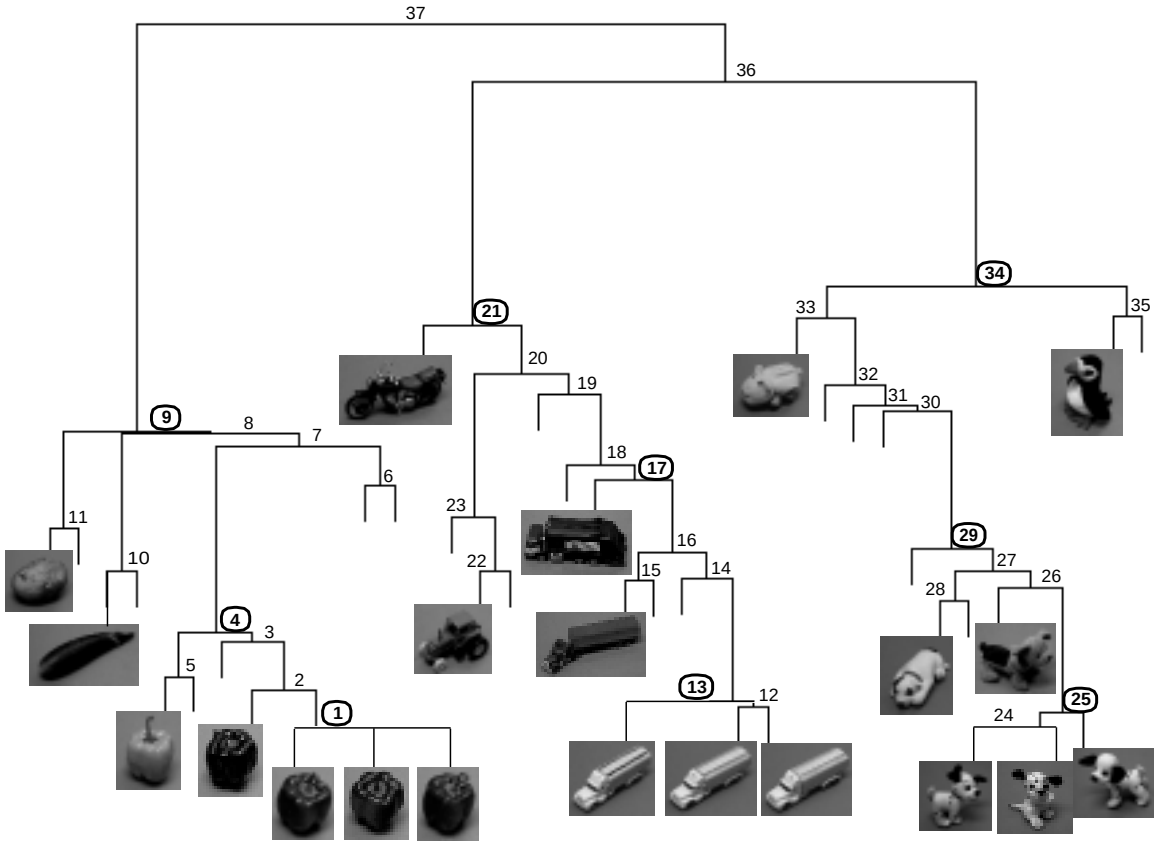



**Figure 7**

corresponding cluster of objects is from the next nearest cluster of objects.

I propose that each node in this tree – each recovered cluster – may be a candidate extension for the novel word (*i.e.* the word would apply to all and only the objects falling under that node in the tree). There are several reasons why this proposal is especially apt for generating hypothetical extensions of object kind terms, particularly *natural* kind terms. First, we have the basic intuition that members of a kind are, on average, more similar to each other than to other objects. This does not mean that any one member of the kind cannot be more similar to a member of a different kind than to some member of its own kind, only that on average this is not true. Average-link clustering captures this intuition by finding clusters whose average within-class similarity would only be increased by the addition of other objects (Duda & Hart, 1973). More deeply, the candidate extensions generated by any hierarchical clustering procedure are guaranteed to be either disjoint or nested, *i.e.* not partially

overlapping, and to be related by inclusion in a tree structure. Tree structures are the standard model for natural category systems in cognitive psychology (Rosch et al., 1980; Murphy & Smith, 1982; Mervis & Crisafi, 1982), and there is some cross-cultural evidence that they are a universal feature of folk biological systems (Coley, Medin, & Atran, 1998; Atran, 1998). Also, the restriction to non-partially-overlapping categories is equivalent to the "M-constraint" of Keil (1979) and Sommers (1963), which Keil argues is a fundamental constraint on peoples' earliest (and deepest) ontological systems. Finally and most relevantly, the assumption that common nouns map onto taxonomic categories appears to be a major guiding principle in word learning (Markman, 1989).

Looking at the tree produced by hierarchical clustering (Figure 7), we can see that it captures in an objective fashion much of our intuitive knowledge about this domain of objects. Each of the nine main classes that we used in designing the stimuli (vegetable, vehical, animal, pepper, truck, dog, green pepper, yellow truck, and dalmatian) corresponds to a node in the tree (marked by a circled number). Moreover, most of these nine clusters are highly distinctive (*i.e.* well-separated from other clusters), as one would expect for the targets of kind terms. A notable exception is the cluster corresponding to trucks (#17), which is barely separated from the next higher cluster. By and large, though, the ideal hypothesis space that we constructed based on intuition (Figure 6) actually turns out to have been a fairly good approximation to the actual structure latent in participants' similarity ratings. However, there are many other hypotheses here as well, capturing more subtle relations between the objects. For example, cluster #18 includes all of the trucks and also the yellow schoolbus. While the schoolbus does not fall into the class of trucks, it comes much closer intuitively than any other non-truck object in the set. Moreover, it is conceivable that a language could have a word that applies to trucks and schoolbuses (and buses), but not other kinds of vehicles like motorcycles and tractors. Other examples of namable nodes include: cluster #36, containing all and only the toys; cluster #23, containing the tractor, the bulldozer, and the crane, but no other vehicles; cluster #33, containing all and only the mammals; etc.

I am not suggesting that hierarchical clustering on similarity is always – or even usually – the right way to generate a hypothesis space of candidate extensions for new words, or more generaly, for new concepts. Even within the domain of natural object kinds, the taxonomic assumption and the M-constraint have their limitations (Carey, 1985). In other domains (including those in both the previous chapter and the next chapter), psychologically natural hypothesis spaces are frequently non-tree-structured. But when we do have reason to believe that a taxonomic hypothesis space is appropriate, then hierarchical clustering on similarity may be an objective and quite powerful way to construct it. In the next chapter, when I consider the need for more complex spaces of overlapping hypotheses, I will briefly discuss how other clustering algorithms (*e.g.* additive clustering) can be used to generate those hypothesis spaces from similarity data as well.

Given the necessary ingredients for rule-based and similarity-based learning algorithms, we now turn to the various models' predictions for our word learning task.

### 4.3.2 Models based on rules

Recall the most basic of rule-based models of concept learning, MIN RULE. MIN RULE chooses the smallest hypothesis $h^*$ consistent with all observed examples and generalizes to all and only those new objects that fall under $h^*$. Several authors in the word learning literature have proposed essentially this algorithm, either explicitly (Berwick, 1985) or implicitly (Clark, 1973). Figure 8 shows the predictions of MIN RULE using the hypothesis space obtained by an average-link clustering of the similarity judgments (Figure 7). Figure 9 shows the predictions of MIN RULE using the ideal hypothesis space of only nine hypotheses (Figure 6). Dark bars show the original data, and light bars show the model predictions. To make the analysis more revealing, the data and predictions for each individual test object are plotted with a separate bar. Because neither the data nor MIN RULE show any generalization outside the superordinate category that the examples belong to, only the eight test objects belonging to the same superordinate are shown in each subplot. These include two subordinate matches (bars 1-2, from left to right), two basic matches (bars 3-4), and

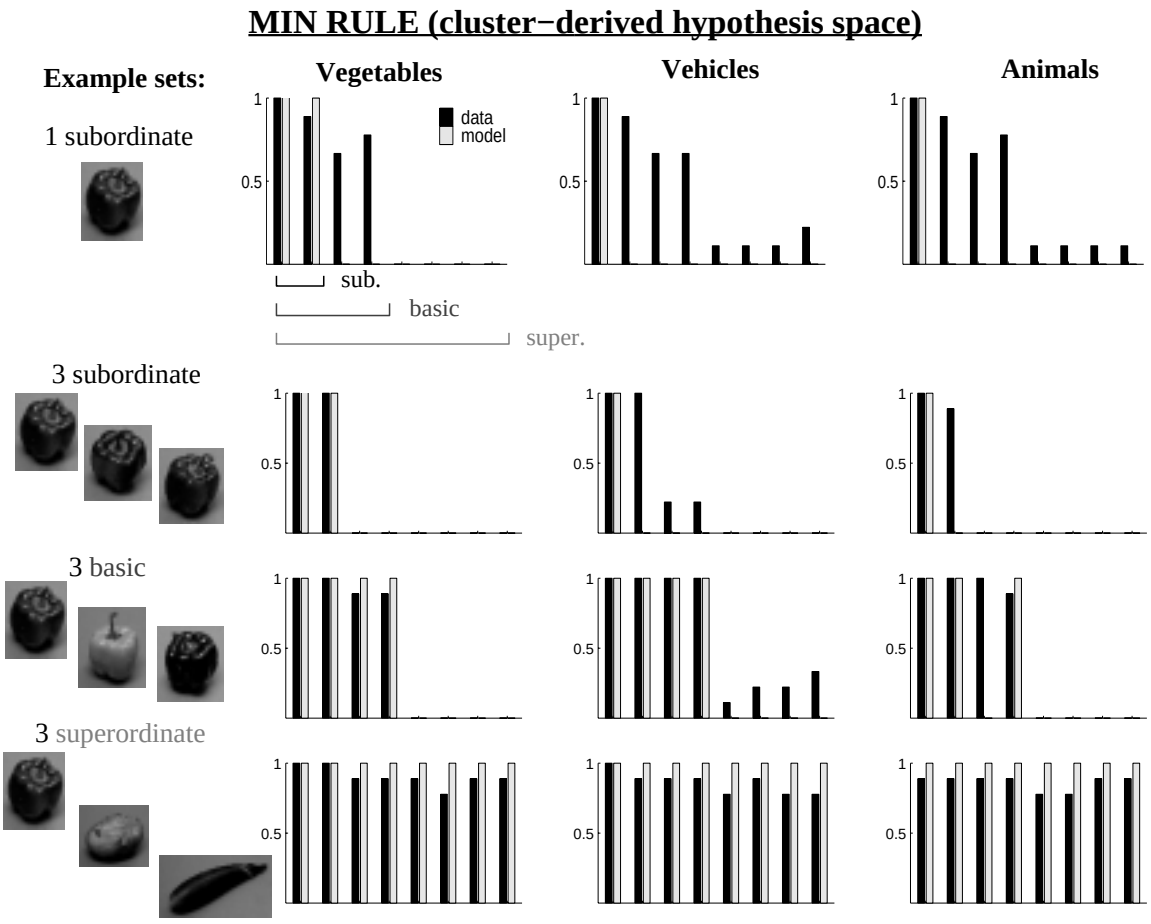



**Figure 8**

four superordinate matches (bars 5-8).

For trials with three labeled examples (rows 2-4), particularly at the basic level and above, MIN RULE provides a satisfactory fit (missing at most one probe stimulus per trial). Using the ideal hypothesis space provides a slightly better fit than using the cluster-derived hypothesis space, because the three labeled examples do not always cover the ideal subordinate, basic, or superordinate category node in the more complex tree of Figure 7. But the difference is minor. In short, on those trials where people generalized in practically an all-or-none fashion, the MIN RULE model provides a reasonable account of their behavior.

However, MIN RULE was not as successful at modeling trials with only one labeled example (row 1). Here generalization falls off in a more graded fashion, so a model that always chooses the most specific hypothesis – or any single hypothesis – cannot

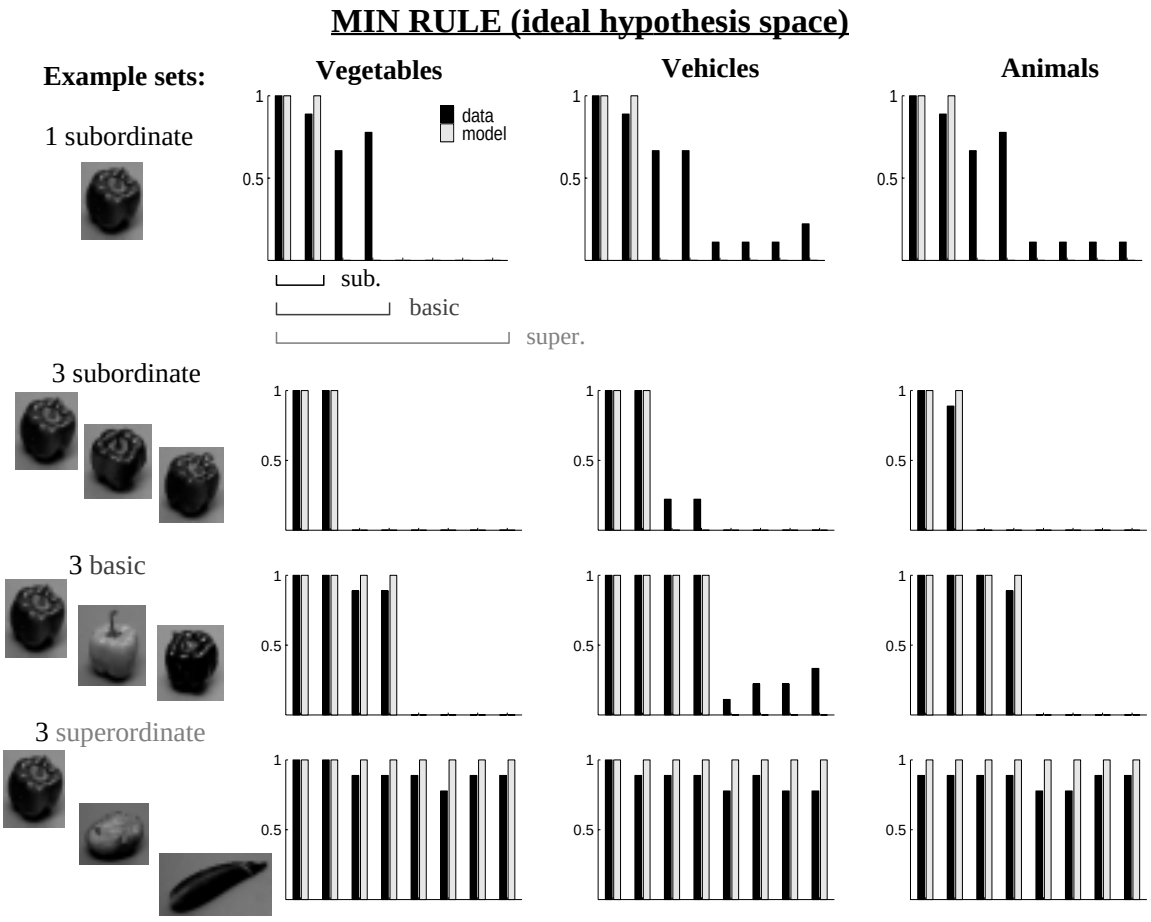



**Figure 9**

very well capture this behavior. Figure 10 sums up the performance of these two models, by plotting the observed frequency of generalization versus the predicted probability of generalization for all test objects over all trials. The fact that data points mostly fall above the diagonal line reflects the conservatism of MIN RULE. Table 1 summarizes the model fits in terms of the correlation between the predictions of each model and the observed frequencies of generalization, measured across all test objects (1st column) and also measured across only those test objects in the same superordinate class as the example objects (2nd column), which is the source of almost all of the variance in participants' data.

Because we only collected yes/no judgments from participants, we cannot conclusively eliminate the possibility that generalization from a single example, despite its graded appearance in the aggregate data, was in fact rule-based on a subject-

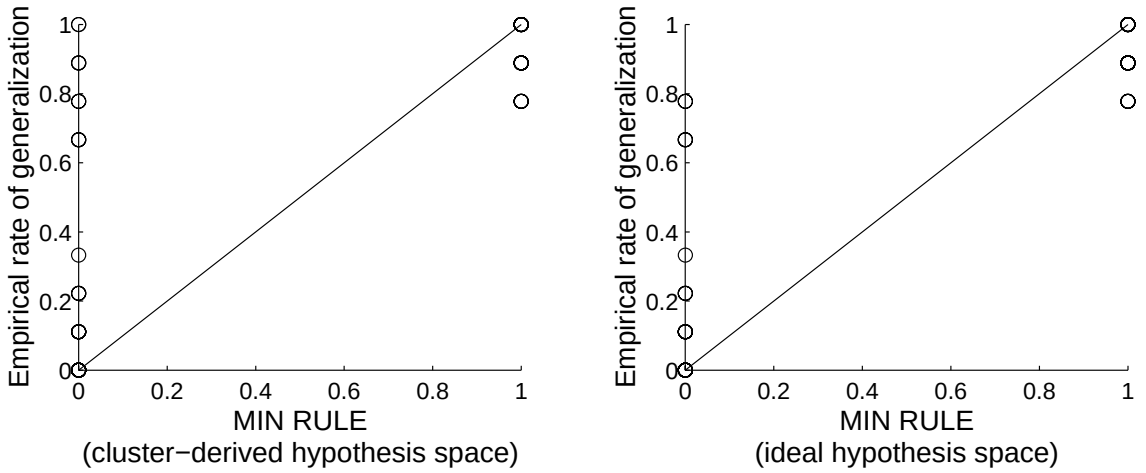



**Figure 10**

by-subject or trial-by-trial case. Suppose that people infer not the single smallest consistent rule, but the single rule that maximizes a combination of small size and high a priori probability. This was called the MAP algorithm at the end of Chapter 2. In that case, if people have different prior probabilities, some could infer from one example (*e.g.* a green pepper) that the word picks out the subordinate class (all and only green peppers) while others could infer that it picks out the basic-level class (all and only peppers). Then when we average over judgments over subjects, generalization would appear to be graded even though each individual subject inferred a single definite rule. It is difficult to rule out this possibility without a more sensitive dependent variable, or a more fine-grained within-subjects analysis than could be supported by our data.

Nonetheless, there is some circumstantial evidence which suggests that individual participants did not infer a single rule from one labeled example as they seemed to do from three labeled examples. This comes from looking at the order in which test objects were picked on individual trials. In particular, we looked at two sets of trials for which people saw different labeled input but made the same generalizations: trials in which a participant selected all four test objects in the same basic-level class (*e.g.* two green peppers and two non-green peppers) after seeing *one* labeled example, and trials in which a participant selected the same four test objects after

| Model | Correlation ($R^2$) with generalization data (all test objects) | Correlation ($R^2$) with generalization data (test objects in same superordinate class) |
|---|---|---|
| Rule-based: | | |
| MIN RULE (cluster hyp. space) | 0.758 | 0.642 |
| MIN RULE (ideal hyp. space) | 0.899 | 0.840 |
| Similarity-based: | | |
| MAX SIM | 0.662 | 0.492 |
| AVG SIM | 0.593 | 0.250 |
| Weak Bayes | 0.730 | 0.801 |
| Strong Bayes (cluster hyp. space) | 0.940 | 0.888 |
| Strong Bayes + basic-level bias | 0.975 | 0.957 |

**Table 1**

seeing *three* examples spanning the basic-level category (*e.g.* a green, an orange, and a purple pepper). [4] In both of these cases, the contents of people's generalizations were equally consistent with either a rule-based strategy – applying a single rule at the basic level – or a similarity-based strategy – responding to all test objects that surpassed a certain level of similarity to the labeled example(s). But the order in which these four test objects were selected was quite different in the two cases. Given only one example (green pepper), the two subordinate (hence more similar) matches (green peppers) were on average chosen significantly earlier than the two basic-level matches (non-green peppers). The average rank (out of 4, with 1 indicating first choice) of the subordinate matches was 1.77, while the average rank of the basic-level matches was 3.23 ($p < .001$, $t = 6.11$, $df = 50$). However, given three examples at the

[4] This response pattern occurred on 13/27 trials with one example and 21/27 trials with three examples that spanned a basic-level class.

basic level (green, orange, and purple pepper), the two subordinate matches were *not* chosen significantly earlier, with an average rank of 2.29 versus 2.71 for the basic-level matches ($p > .08$, $t = 1.77$, $df = 82$). Thus we can conclude that people were doing something quite different in these two cases, even though their generalization choices were identical. Given three examples, people seemed to infer that the word referred to the basic-level class and treated all probe objects in that class equivalently, picking them out in more or less random order. Given one example, people seemed to infer that the word definitely applied to the highly similar subordinate matches, and, after some reflection, that it might also apply to the less similar basic-level matches as well. MIN, MAP, or any other model in which people generalize according to a single best rule has no way to account for this preference gradient observed on single-example trials.

### 4.3.3 Models based on similarity

The basic similarity-based model we introduced in the last chapter, MAX SIM, assumes that the probability of generalizing to a test object $y$ is proportional to the maximum pairwise similarity of $y$ to each of the labeled examples. Figure 11 shows the predictions of MAX SIM using the pairwise similarities among objects as rated by participants in the experiment. With only one labeled example (row 1), MAX SIM captures the graded nature of people's generalization – exactly what MIN RULE missed. However, the MAX SIM gradient does not fall nearly as fast as people's generalization does. This causes MAX SIM to fit particularly poorly with three labeled examples (rows 2-4), where people are generalizing in a practically all-or-none manner. Figure 14a shows that globally, the predictions of MAX SIM are very only weakly correlated with the observed data (see also Table 1).

To show that these poor results are not simply an artifact of the particular way in which MAX SIM computes similarity to the set of examples, we also tested the predictions of a different similarity-based model, AVG SIM. AVG SIM defines the probability of generalizing from a set of exemplars $X = \{x_1, \ldots x_n\}$ to a new object

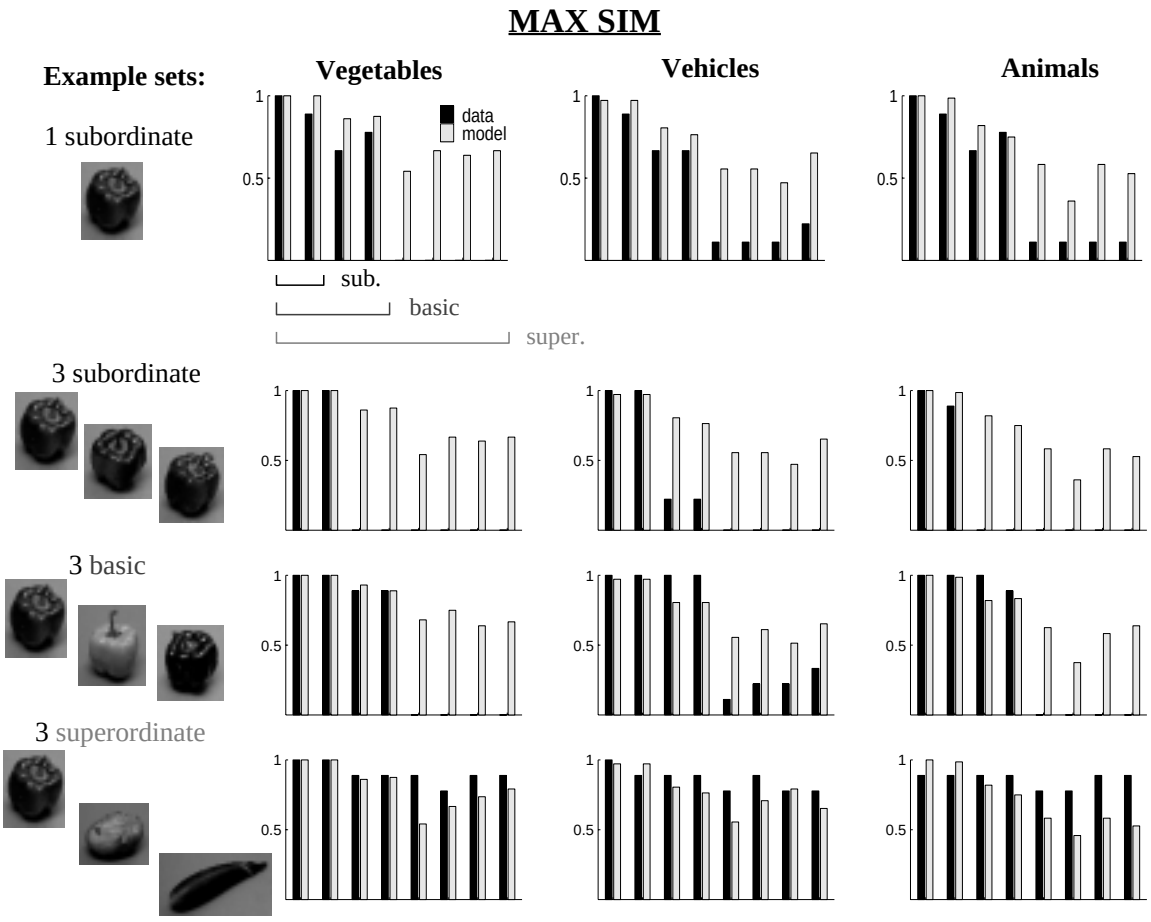



**Figure 11**

$y$ to be the *average* pairwise similarity of $y$ to each of the examples $x_i$,

$$p(y \in C|X) = \frac{1}{n} \sum_i \text{SIM}(y \rightarrow x_i). \tag{4.1}$$

AVG SIM is in the same spirit as the classic exemplar models of classification, Medin & Schaffer's (1978) *context model* and Nosofsky's (1986) *generalized context model*. However, AVG SIM is not the answer here. Figure 12 shows the specific predictions of AVG SIM, and Figure 14b shows the global relationship between AVG SIM's predictions and the observed data. In terms of correlation with the observed frequencies of generalizaion, AVG SIMperforms even worse than MAX SIM (Table 1).

In Chapters 2 and 3, I argued that a variant of the Bayesian framework, Weak Bayes, instantiates a principled similarity-based approach to concept learning. Fig-

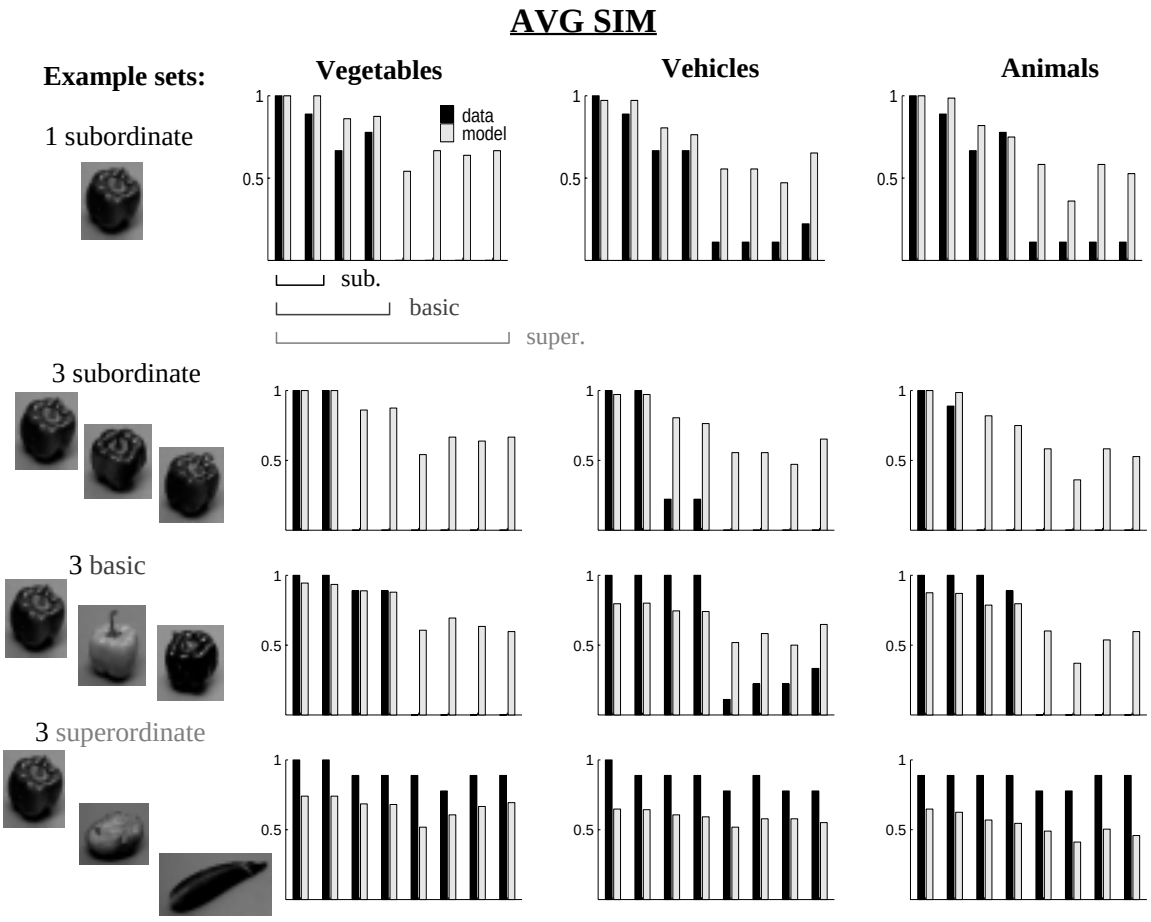



**Figure 12**

ures 13 and 14c show the predictions of Weak Bayes for this task. [5] These predictions are qualitatively similar to but somewhat more accurate than the predictions of AVG SIM and MAX SIMÅs the correlations in Table 1 show, Weak Bayes gives a particularly better account of how participants generalized from three labeled examples to other objects in the same superordinate class, and somewhat less accurate a picture of generalization behavior across the entire set of test objects. Still, it never achieves a value of $R^2$ better than 0.8.

Although models based on similarity to exemplars have proven successful at accounting for patterns of classification learning behavior using artificial stimuli (Nosof-

[5]The details of the Bayesian framework are described below. Weak Bayes is equivalent to the Strong Bayes model presented in Figures 16 and 18a, except with a likelihood term that measures only whether a hypothesis is consistent with the observed examples ($p(X|h) = 1$ if $h$ contains the examples $X$ and 0 otherwise).

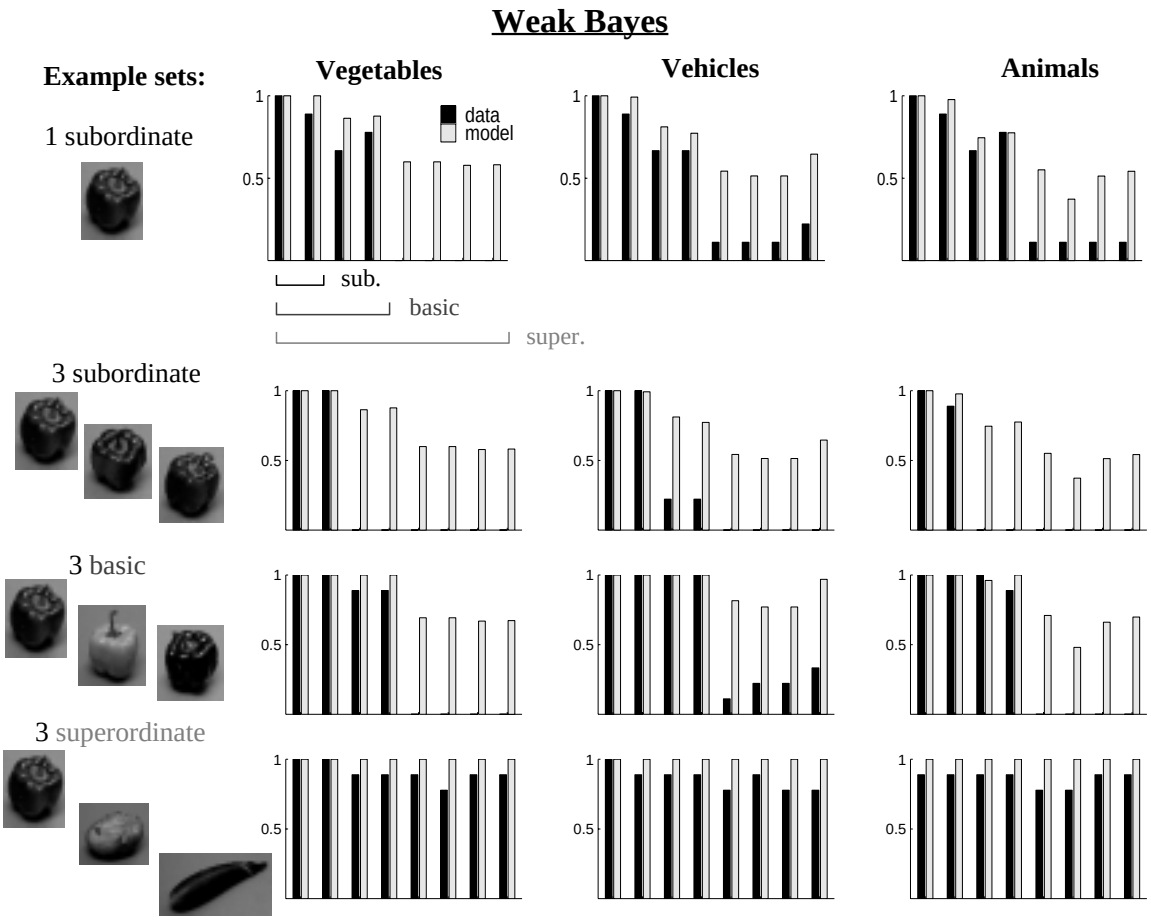



**Figure 13**

sky, 1992), they do not seem directly applicable to problems of concept learning from several positive examples, like the one we face here. Figures 11 and 12 reveal the heart of the problem. Compare the case of one labeled example, *e.g.* a green pepper, versus three labeled examples all from the same subordinate class, *e.g.* three green peppers. A test object from another subordinate in the same basic class, *e.g.* a red pepper, is not significantly more or less similar to each of the three green pepper examples than to the single green pepper example; *i.e.* average or maximum similarity to examples does not change for this test object with the addition of the two more labeled examples. [6] Yet the probability of generalization goes way down, as people look onto the

[6]The third similarity-based approach introduced in Chapter 1, "total similarity", is not even appropriate for modeling generalization frequencies at all, because its predictions of generalization are not normalized to lie between 0 and 1. They could be normalized by dividing them by the number of labeled exemplars, but that is exactly equivalent to AVG SIM.

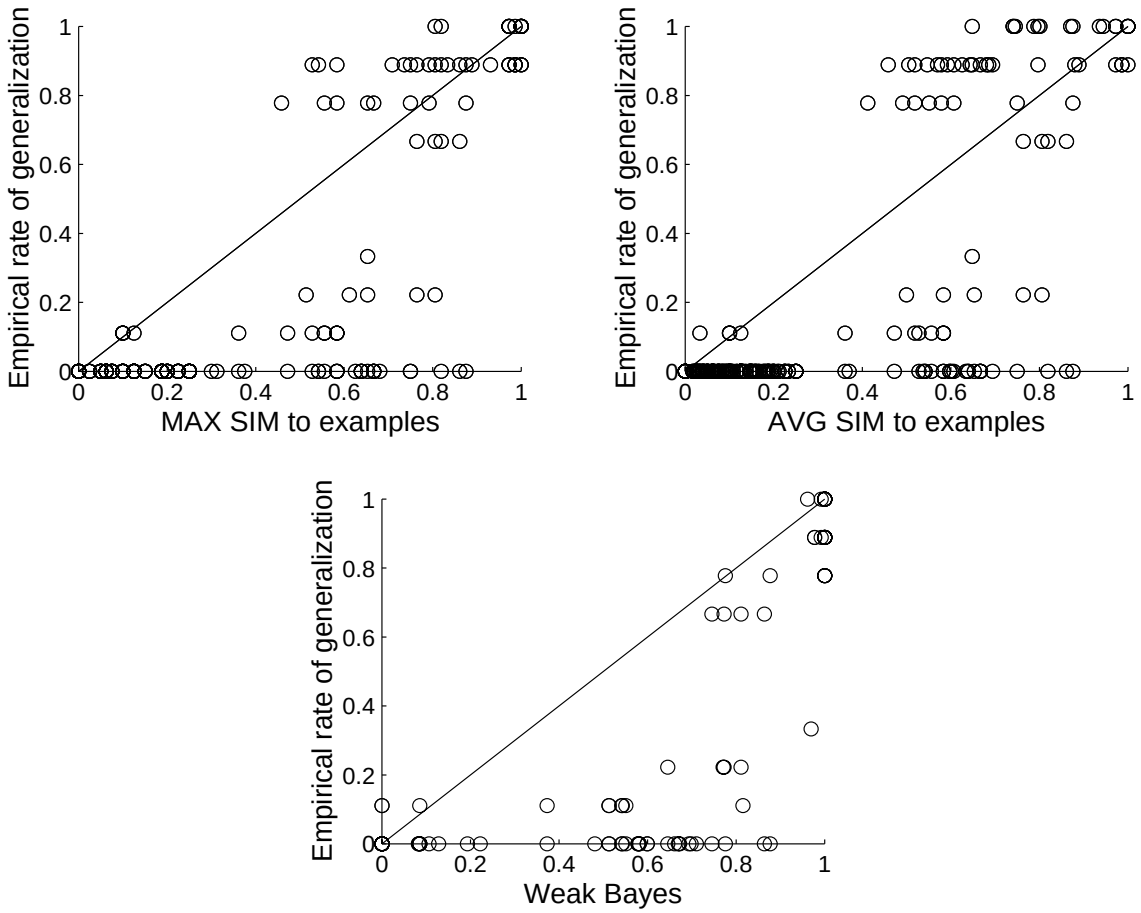



**Figure 14**

most specific hypothesis containing the three examples, *i.e.* green peppers. Failing to capture convergence to a highly specific, all-or-none state of generalization is the main limitation of MAX SIM AVG SIM and Weak Bayes – and is precisely where the complementary approach of MIN RULE excelled.

### 4.3.4 The Strong Bayes model

By now, if not long before, this story should be starting to sound familiar. The basic phenomena of generalization we obtained, and the ways in which they frustrate traditional concept learning algorithms based strictly on rules or similarity, parallel quite closely what we found in the healthy levels case study. In both cases, initially broad gradients of generalization converge, as more examples are observed, to the

most specific rule consistent with all the examples. The same principles that allowed the Strong Bayes model to explain this transition from similarity-like to rule-like generalization on the healthy levels task also support an explanation of the analogous transition on the word learning task.

Again, we list the four ingredients of Strong Bayes:

1. A constrained **hypothesis space** of possible extensions of a concept;

2. A **prior distribution** over the hypothesis space reflecting the learner's relevant background knowledge;

3. The **size principle** for scoring the likelihood of hypotheses under the *strong sampling* generative model, favoring smaller consistent hypotheses with exponentially greater weight as the number of observed examples increases;

4. The notion of **hypothesis averaging**, *i.e.* integrating the predictions of multiple consistent hypotheses, weighted by their posterior probabilities, to arrive at the probability of generalizing a concept to a new entity.

The hypothesis space – ingredient 1 – is already given by the hierarchical clustering solution we developed for the simple rule-based approaches (Figure 7). It remains for us to determine ingredients 2 and 3, appropriate prior and likelihood terms for scoring these hypotheses based on their a priori plausibility and their probability of generating the observed data. Fortunately, the average-link clustering model of similarity judgments already contains both of these kinds of information implicitly represented in the taxonomic tree it generates.

First consider the likelihood term. Recall the content of the *size principle* underlying Strong Bayes: the probability of observing $n$ examples consistent with hypothesis $h$ is inversely proportional to the size of $h$ raised to the $n$th power. This principle follows as a simple consequence of assuming that examples are sampled from a uniform distribution over $h$. Thus what we really need is a measure of the *size* of each hypothesis in our hypothesis space. Now, there is an intuitive (negative) relationship between the size of a class and the average similarity of its members to each other:

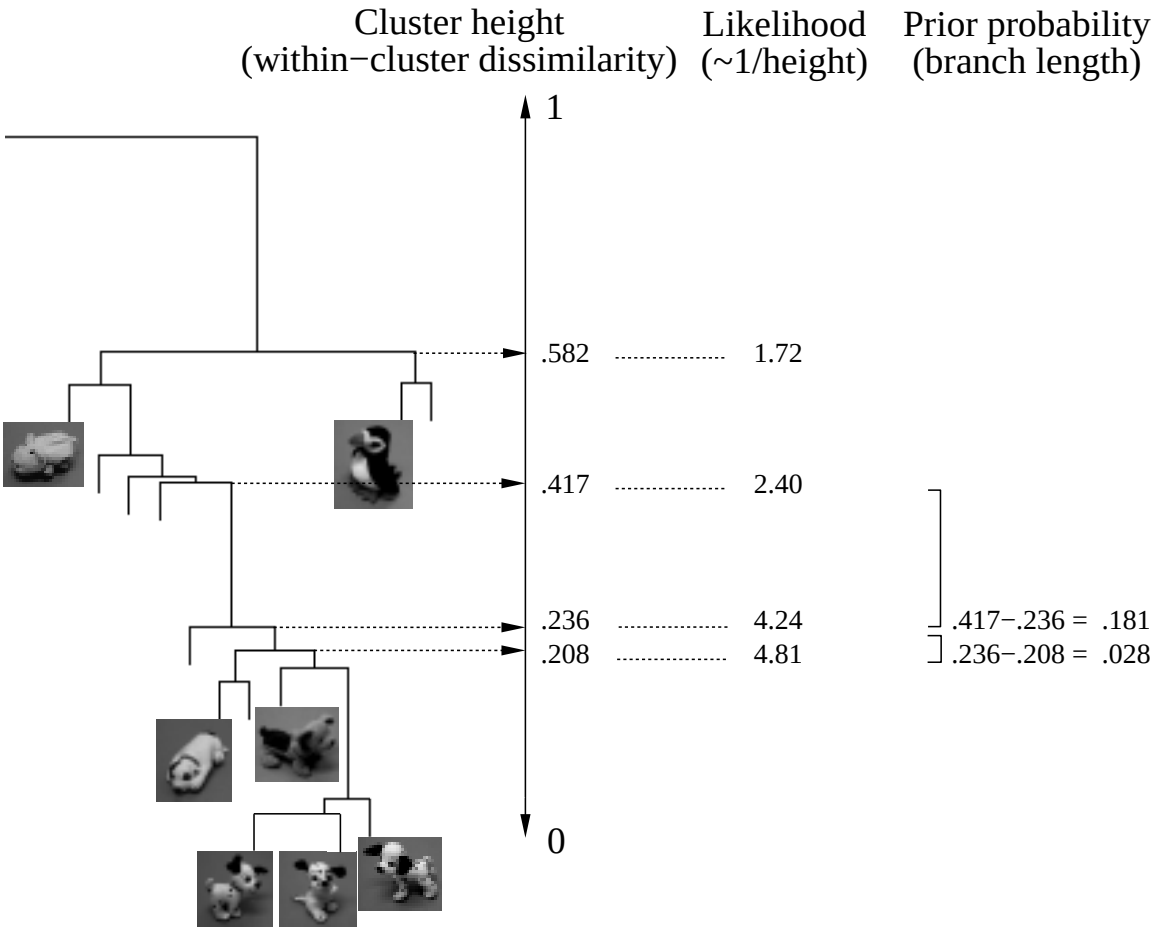



**Figure 15**

the smaller the class, the higher the average within-class similarity, and vice versa. So the simplest measure of the "size" of each hypothesis is just one minus its average similarity. I put size in quotes because this isn't the physical size of any real set; it's the subjective size of a candidate extension, as reflected in how (dis)similar people think its members are to one another on average. Note that this measure of a hypothesis' size is identical to the height of the corresponding node in the tree of Figure 7, *i.e.* lower nodes in the tree have higher average within-class similarity. Thus we take the likelihood of hypothesis $h$, given the examples $X = \{x_1, \ldots, x_n\}$, to be to be

$$p(X|h) = \left[\frac{1}{\text{height}(h)}\right]^n \tag{4.2}$$

if $x_i \in h$ for all $i$, and 0 otherwise. Figure 15

illustrates the relation between the likelihood of a hypothesis and the height of the corresponding node in the tree produced by hierarchical clustering. In practice, we must add a small constant $\epsilon > 0$ to height($h$), or else the likelihood term blows up for the lowest nodes in the tree (which have height 0). The exact value of $\epsilon$ is not critical; we found the best results with $\epsilon = 0.05$.

Next consider the prior. Intuitively, a class whose members are much more similar to each other than to objects outside the class seems like a good candidate concept, an a priori natural hypothesis for the extension of a new word. In contrast, a class whose members are no more similar to each other than to objects outside the class seems like just another random collection of objects, not worthy of name or note. Earlier, I used the term "distinctiveness" to refer to the average similarity of objects within a class relative to the nearest objects outside that class – essentially a local version of the intuition behind what makes a good candidate concept. Also, I noted that the distinctiveness of a cluster was represented in tree of Figure 7 by the length of the branch above the corresponding node. For example, the class containing all and only the dogs (#29) is highly distinctive, but the classes immediately under it (#27) or above it (#30) are not nearly as distinctive. Thus we take the prior probability to be

$$p(h) = \text{height}(PARENT(h)) - \text{height}(h), \tag{4.3}$$

where $PARENT(h)$ denotes the parent node of $h$ in the tree. Figure 15 illustrates this branch-length prior, which gives a roughly 9-to-1 preference for the class containing all and only the dogs (#29) versus the class immediately under it (#27), which contains all dogs except one. The role that this prior probability plays in guiding generalization is directly analogous to its role in Chapter 2, where, in explaining why we generalize from 16, 8, 2, 64, to all powers of two, it was necessary to assign the hypothesis *all powers of two except 32* very low prior probability to make up for its slight advantage in likelihood over the hypothesis *all powers of two*.

Figure 16 shows the predictions of Strong Bayes when all of these ingredients (together with hypothesis averaging) are combined to compute the probability of

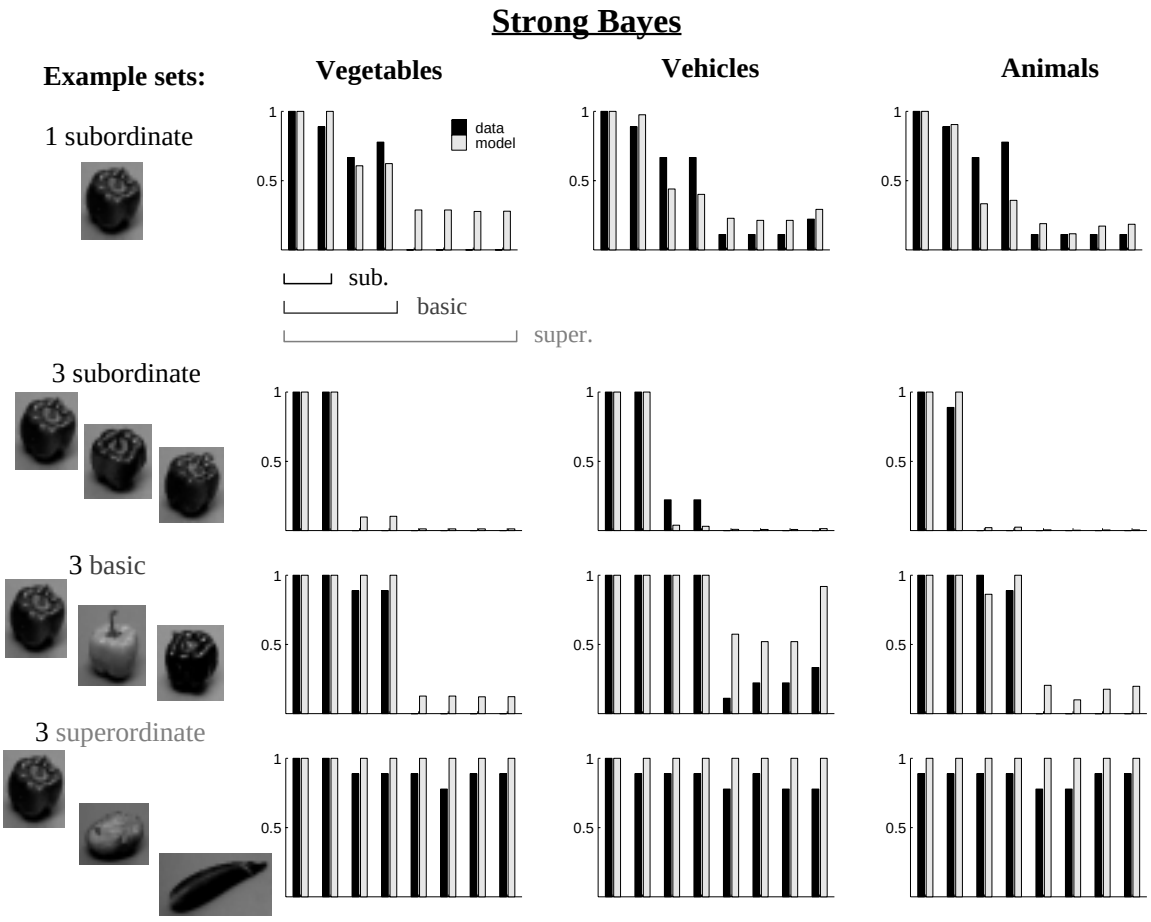



**Figure 16**

generalization to a new object $y$:

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X). \tag{4.4}$$

In contrast to the strictly rule- or similarity-based approaches considered above, the Bayesian model generalizes more or less in a graded fashion after one example, and in an all-or-none fashion after three examples. Figure 18a shows the global correlation of the strong Bayes model with the observed data. The variance accounted for (Table 1) is clearly an improvement over the strict rule-based or similarity-based models, but it is still far from perfect. Looking back at Figure 16, we see a few glaring errors that could account for the discrepancy. We will come back to those errors shortly.

The intuition for why the model switches from graded to sharp generalization

here is the same as in previous chapters: the combination of hypothesis averaging and the size principle. Hypothesis averaging means that the more hypotheses that include both the labeled examples and a test object $y$, and the more probable those hypotheses are, the greater the probability of generalizing to $y$. The size principle has the effect of weighting smaller hypotheses as more likely than larger hypotheses, just as in the number concept or healthy levels tasks. Just as it would seem very unlikely to observe the numbers 16, 8, 2, and 64 if we were sampling from all even numbers or all numbers under 100, as opposed to only the powers of two, so it seems very unlikely to observe three green peppers if our examples were drawn randomly from the set of all peppers or the set of all vegetables, as opposed to just the set of green peppers. This preference is more extreme as we observe more examples; with just one green pepper, it seems like hardly a coincidence at all.

Hypothesis averaging and the size principle interact as follows to explain the similarity-to-rule transition in generalization. From any one example, *e.g.* a green pepper, there are many possible ways to generalize: to all and only green peppers, all and only peppers, all and only vegetables, not to mention other classes in the tree that do not have common names in English. Different hypotheses receive different probabilities based on a combination of prior and likelihood terms; the likelihood prefers smaller over larger hypotheses, but these preferences are relatively weak after only a single example. Probability of generalization is computed by averaging the predictions of all hypotheses weighted by their probabilities. Because no single hypothesis dominates the others in probability, the resulting generalization function appears fuzzy – reflecting the conflicting predictions of multiple rules – and graded – as a function of how many candidate rules a test object satisfies. On the other hand, after three examples are observed, *e.g.* three green peppers, the size principle's preference for smaller hypotheses is multiplied exponentially, and the smallest consistent hypothesis (all and only green peppers) takes on a much greater likelihood than any of the alternatives. Now, following the weighted average of all hypotheses is essentially equivalent to following only the single most likely and the Bayesian learner locks in on the appropriate rule for generalizing the new word.

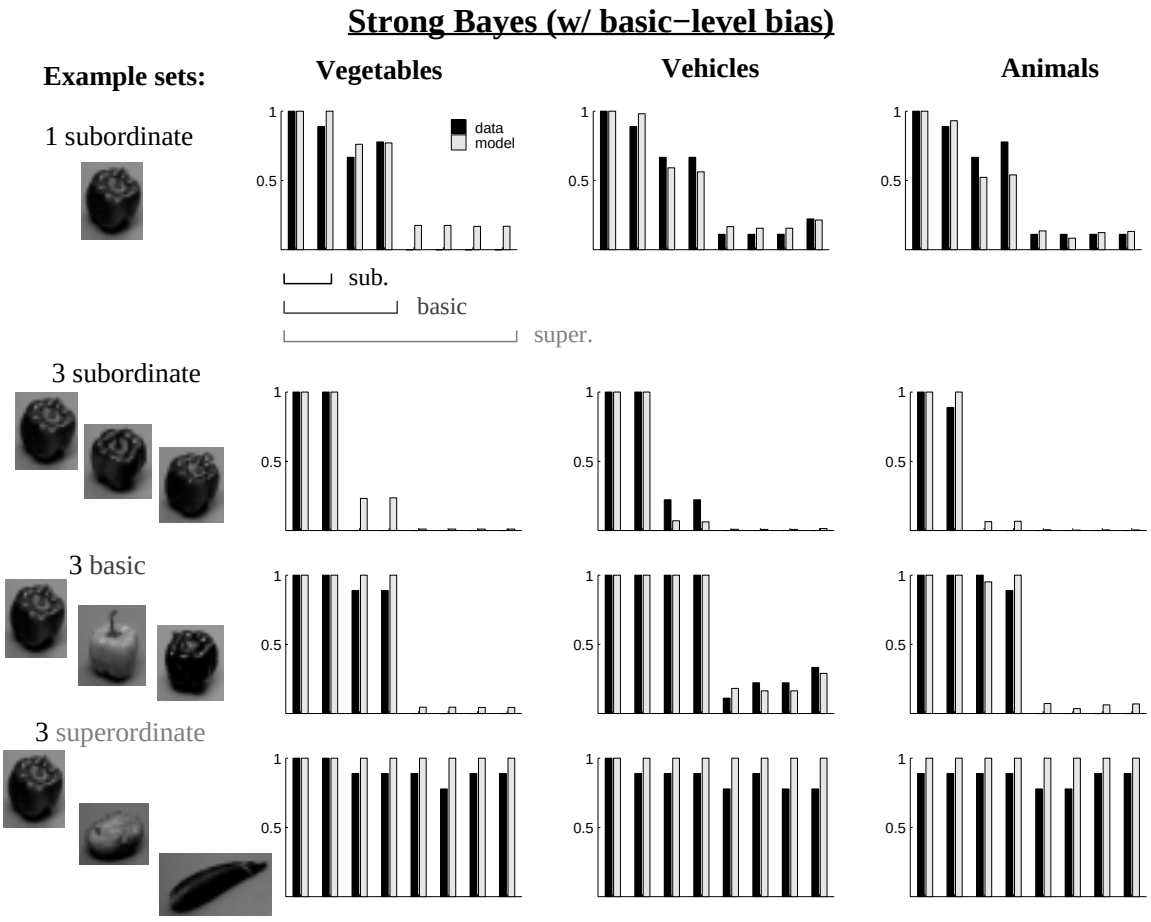



**Figure 17**

Returning to Figure 16, we can see two main places where the Bayesian model diverges significantly from people's generalization behavior. First, given one example (row 1), the model does not generalize enough to test objects matching at the basic level. Second, given three examples in the same basic-level class (row 3), the model generalizes *too much* to test objects outside that class (particularly columns 2 and 3). Together, these two kinds of errors suggest that the model does not have as strong a preference as people do to map words onto basic-level categories.

Earlier in this chapter, we mentioned several reasons why people might have such a basic-level bias on this task. Without necessarily taking a stand on which of those, if any, is correct, we can easily incorporate a basic-level bias into the model and see whether that significantly improves the fit. To do so, we added a component to the prior probability of each of the three hypotheses corresponding to our participants'

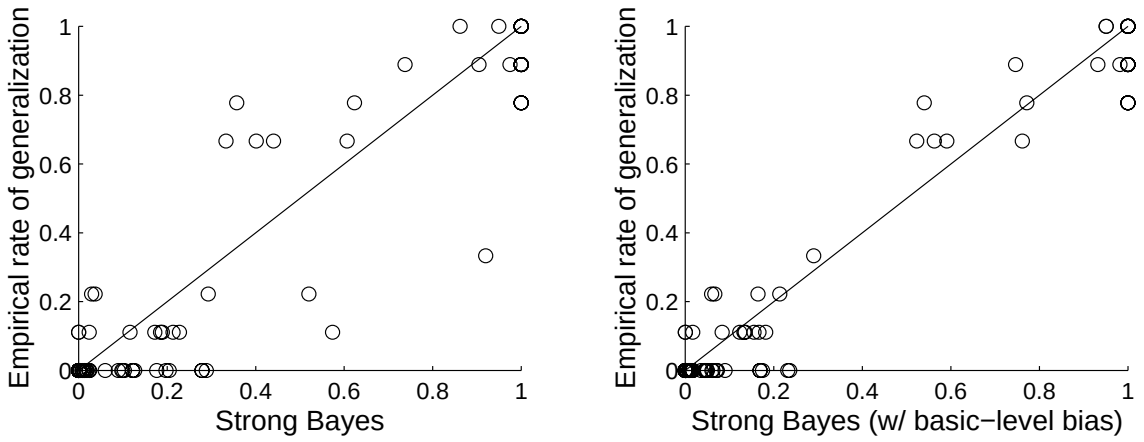



**Figure 18**

presumed basic-level categories: "pepper", "truck", and "dog". [7] Everything else about the model was unchanged. Figure 17 shows that with the addition of this one free parameter, the predictions for individual test objects become much closer to the empirical frequencies. Figure 18b and Table 1 show that the global correlation between model and data is now almost perfect, with $R^2$ values greater than 0.95.

## 4.4 Discussion

The abstract problem of this thesis, concept learning from one or a few positive examples, finds one of its most important concrete expressions in the task of word learning. Likewise, the foundational contemporary work on word learning places the child's problems of induction and generalization at center stage (Quine, 1960; Carey, 1978; Macnamara, 1982; Markman, 1989). The great debates that have framed the study of concept learning and induction more generally – Is acquisition driven mainly by pre-existing knowledge or the statistical force of our observations? Is generalization based primarily on rule-like representations or similarity to exemplars? – are also among the central questions of word learning (Bloom, in press; Carey, 1982; Jones & Smith, 1993). In this chapter, I have tried to argue that a formal Bayesian

[7] This component was treated as a free parameter. Here, it was set to equal an additional branch length (recall that the prior probability of each hypothesis is proportional to the length of the corresponding node's branch in the tree of Figure 7) of half the total tree height.

framework for concept learning allows us to view these issues in word learning in new and productive ways.

First, I presented empirical evidence that the way people generalize words from one or more labeled examples appears to have both rule-based and similarity-based aspects to it. This interpretation was supported by showing that simple learning algorithms which incorporate only rules or similarity are not capable of modeling our data. As in the more abstract concept learning tasks studied in the preceeding chapters, similarity-based models are most applicable for modeling the broad gradients of generalization observed in the earliest stages of learning, while rule-based models become more reasonable as additional examples are encountered and learners restrict generalization to the most specific natural class containing all the examples.

I then presented a Bayesian analysis of word learning, which treats the labeled examples of a word as data and the problem of generalizing the word to new objects as Bayesian inference over a hypothesis space of candidate word extensions. The Bayesian model generates both similarity-like and rule-like behavior in the appropriate regimes. We explained the transition between these two modes of generalization in the Bayesian framework in terms of the *size principle*, which causes an initially broad posterior probability distribution to become strongly concentrated on the smallest consistent hypothesis after just a few examples, and *hypothesis averaging*, which leads to graded or sharp generalization behavior depending on whether the posterior probability distribution is broad or sharply peaked.

In the context of the debate between knowledge-driven and data-driven theories of word learning (Bloom, in press), the present work clearly shows the need for both factors. The Bayesian framework gives prior knowledge an explicit role in fixing the hypothesis space of candidate extensions and assigning a prior probability over that space. Constraints on word learning – the mainstay of the knowledge-driven approach – naturally enter into this stage of Bayesian modeling. In particular, we incorporated two well-known constraints from the literature on children's word learning: the tree-structured hypothesis space was suggested by the taxonomic bias (Markman, 1989), while the prior probability embodied a bias for mapping words onto basic-level cate-

gories (Markman, 1989; Mervis & Crisafi, 1982). However, all of this prior knowledge cannot on its own account for how participants in our study generalized, without also acknowledging the statistical role of the input data. Given a nested hierarchy of candidate word extensions, there will typically be many *a priori* natural ways to generalize a given set of examples; given three green peppers as examples of "blicketts", "blicket" could refer to all and only green peppers, to all and only peppers, to all and only vegetables, etc. We need to invoke the size principle – fundamentally, a statistical constraint deriving from a probabilistic model of the observed examples – to explain how people are able to lock in on the true extension of a word after seeing only a few labeled examples. The Bayesian framework thus provides a formal model of how prior knowledge – including conventional word learning constraints – interacts with the word learner's input data to guide generalizations of word meaning from very limited evidence. Neither data nor priors are much help without the other. Indeed, the usual debate (Bloom, in press) about whether word learning is possible in virtue of powerful a priori knowledge, or powerful statistical algorithms, doesn't even make sense in a Bayesian context. For the Bayesian learner, the observed data have great statistical power, but they have that power precisely because of the strong models that the learner brings to the task.

To illustrate how radically different this Bayesian framework is from what is often thought of as statistical learning in the word learning literature, consider the distinction Bloom (in press) draws between empiricist and rationalist theories of word learning. Empiricist theories claim that acquiring word meanings depends fundamentally on noticing correlations between features in the environment, while rationalist theories claim a central role for the learner's knowledge about possible individuals, speaker's intentionality, and other such abstract theoretical notions. In Bloom's view, and probably in the eyes of many people on both sides of the debate, statistical learning mechansims are primarily the province of empiricist or neo-empiricist (*i.e.* connectionist) theories. However, rationalist notions like "individual" and "intentionality" are actually essential pieces of the Bayesian framework presented here. A Bayesian learner needs to know how many distinct examples of a concept (*i.e.* how

many independent pieces of evidence) she has seen, in other words, what $n$ should be in the expression for the likelihood, $p(X|h) = 1/|h|^n$. This is critical because the likelihood depends exponentially on $n$. In terms of the statistical inferences a word learner can make, there is a big difference between seeing three different green peppers called "blicket" and seeing the same green pepper called "blicket" on three different occasions. Thus the abstract notion of "individual" has a central place in Bayesian concept learning. Barsalou, Huttenlocher & Lamberts (1998) recently presented evidence that subjects in artificial category learning tasks do in fact track individual objects across stimulus presentations, as the Bayesian theory requires.

The notion of "intentionality" also figures into the size-based likelihood. The size principle comes from assuming a particular generative model, or process producing the examples we observe. This is the *strong sampling* model, in which the examples are sampled randomly *from the true extension of the concept.* Without such an assumption, the Bayesian learner has no basis for asserting that three green peppers were more likely to have been sampled from the set of all and only green peppers than from the set of all vegetables. In the context of learning words from how they are used by other people, the strong sampling model is clearly a claim about speakers' intentionality. It amounts to assuming that people do not indiscriminately label all objects in the world according to whether they are positive or negative examples of a word, but that they use a word when they intend to pick out something from a particular subset of the world, *i.e.* that word's extension, and so that labeled object may be treated as having been sampled *from that subset.* Under weaker sampling situations, as when examples are provided in the form of feedback on the learner's own misidentifications, this intentional connection between word and object is no longer there, and a different probabilistic model (*e.g.* Weak Bayes, instead of Strong Bayes) is appropriate. In short, the Bayesian learner does not treat all observed pairings of word $w$ with object-percept $p$ alike – as one might if one were merely collecting feature co-occurence statistics. Rather, the Bayesian is prepared to make very different inferences depending upon what she knows or assumes to be the process generating these observations: *i.e.* whether or not $p$ reflects a random sample of things in $w$'s ex-

tension, whether or not this sample is independent from a different percept $p'$, and so on. Gricean notions, like the assumption that speaker's utterances are "informative" which Bloom has suggested may actually underlie the mutual exclusivity constraint in word learning, are clearly of a similar nature and could be naturally incorporated into the Bayesian framework.

Finally, in light of the rules vs. similarity debate we have considered throughout the thesis, an intriguing aspect of this chapter's Bayesian analysis was the origin of the model's hypothesis space. We found that a reasonable set of candidate rules for word extensions could be obtained from a hierarchical clustering of people's similarity ratings. The hypothesis space produced by hierarchical clustering is guaranteed to have the tree structure of natural kind taxonomies thought to underlie the learning of common object terms (Markman, 1989); the individual clusters are also good candidates for kind term extensions, in the sense of Quine (1969), because their members are on average more similar to each other than to members of nearby clusters. The metric properties of the clusters' average similarity levels were also used to define the prior probability and likelihoods of the Bayesian model. In a deep sense, then, the spirits of both rules and similarity are at work here. Without the metric properties of similarity telling us the relative sizes of different classes, we could not have explained why people's generalization converges after just three examples to the most specific possible rule. But without a hypothesis space of possible rules, we could not have explained how generalization could converge to a rule at all! And without either a primitive measure of similarity or the notion of a rule-based taxonomy, there wouldn't even have *been* a hypothesis space.

### 4.4.1 The Bayesian framework as a model of "flexible similarity"

Committed similarity theorists may want to view the Bayesian framework differently, as an instance of the "flexible similarity" approach to concept learning. Advocates of this position (Goldstone, 1994; Medin, Goldstone, & Gentner, 1993; Nosofsky, 1986;

Jones & Smith, 1993) argue that, despite the evidence against simple similarity-based models like MAX SIM, our primitive sense of similarity is in fact powerful enough to provide the "groundwork for cognition", once we understand that similarity can be flexible, context-sensitive, and computationally sophisticated. The danger in letting similarity be all of these wonderful things is that we lose the one characteristic that made a primitive sense of similarity such an appealing theoretical foundation in the first place – namely, that it was *primitive*. Once anything is allowed to influence and warp similarity, anything will, and the explanatory construct of similarity quickly becomes "an impostor, a quack", to quote Goodman's (1972) famous appraisal.

The only way to prove that flexible similarity is not an impostor is to unmask its true identity, *i.e.* to build a formal model. The mathematical models of Nosofsky (1986; 1992) and Kruschke (1992) have in fact made good on the promise of flexible similarity as the basis for human classification learning in abstract artificial domains. Nosofsky's (1986) groundbreaking idea was to start by constructing a feature space representation of his stimuli (much like the two dimensional space of healthy levels in Chapter 3) using multidimensional scaling (MDS) on similarity data, and then to embed this stimulus representation in a flexible model of classification learning, Medin & Schaffer's (1978) context model. Kruschke (1992) then showed how the flexible parameters of the model (attentional weights, etc.) could be learned to best discriminate a set of positive and negative training instances of a concept. However, these models have yet to address the biggest challenges in making flexible similarity work for more natural tasks, such as learning words for natural object kinds. Two key challenges are: (1) developing an algorithm appropriate for learning concepts from just one or a few positive examples only, and (2) developing representations suitable for dealing with natural stimuli like peppers, trucks, and dogs.

In the paradigm of flexible similarity, the Bayesian framework can be seen as a generalization of the Nosofsky/Kruschke approach to concept learning which is designed to address precisely these two challenges of algorithm and representation. First, instead of using a classification algorithm that can only learn to discriminate positive from negative instances (see Appendix A for a discussion), I adopt the more

powerful machinery of Bayesian inference. The Bayesian learner is quite capable of acquiring concepts from both positive and negative examples, but can also, under a suitable generative model of the examples, infer how to generalize from a small number of positive examples only. Second, instead of using a continuous feature space representation, I construct a hypothesis space of candidate extensions of the concept. As we saw in Chapter 3 with the healthy levels task, the hypothesis space approach is perfectly capable of modeling generalization in continuous feature spaces. [8] Indeed, the Bayesian framework there predicted how gradients of generalization in feature space will shrink or stretch depending on the number and distribution of examples observed, a phenomenon for which the Nosofsky/Kruschke approach cannot formally account without negative examples. However, as I've shown in this chapter, the hypothesis space representation is far more general, including as another important special case a taxonomic hierarchy of object kind classes. Taxonomic hypothesis spaces, just like continuous feature space representations, may be constructed by scaling similarity data (using hierarchical clustering instead of MDS). Hypothesis spaces with yet more complex structure can be constructed with more elaborate scaling procedures such as additive clustering (Arabie & Shepard, 1979; Tenenbaum, 1996; see next chapter for an application).

So, what is the verdict for similarity? Can it be, in some suitably flexible form, the basis for concept learning? While primitive similarity could not (through simple models like MAX SIM or AVG SIM) directly account for how people generalized new words in this study, it did prove to be an important ingredient in the more successful Bayesian model, as a means by which to generate the hypothesis space and to measure the size of hypotheses. Also, in the healthy levels task of Chapter 3, we implicitly invoked a primitive similarity metric in an analogous role when we defined the size of each rectangle hypothesis. Together, these two case studies do bring the construct of flexible similarity closer to computational respectability. However, they hardly show

[8] Tversky (1977) and Tversky & Gati (1982) make an analogous point about the virtues of set-theoretic models of similarity over metric models, which can be treated as a special case within the set-theoretic framework.

that similarity carries the main burden of concept learning. Similarity can only be related to generalization behavior after being processed in a substantial and theoretically motivated way, via the machinery of Bayesian inference. Other ingredients, such as a taxonomic constraint on the hypothesis space, are essentially rule-based. And in some cases, such as the number concept task, a primitive sense of similarity may play little or no role. Once we have a formal system like the Bayesian framework, it no longer makes sense to ask whether similarity – or any other one factor – is "the basis for concept learning". The only meaningful questions concern how, in a particular case, does similarity interact with the other pieces of the framework to generate the behavior that we observe in people or would like to observe in our computers.

### 4.4.2 The relevance for studies of word learning in childhood

Undoubtedly, we must exercise caution in generalizing findings and models of word learning in adults to parallel claims about children. Children have different cognitive resources, and certainly different resources for learning language than adults do. However, it appears that word learning is very different from syntax acquisition, in that it does draw on domain-general learning mechanisms which adults and children have in common for the most part (Markson & Bloom, 1997; Bloom & Markson, 1998). Hence studying adults could tell us a lot about the general strategies involved in children's word learning, if not the specific details.

One major difference is that children do not already possess a whole vocabulary of English words – as our adult learners did – when they begin to learn words. This could explain why generalization in our experiment was much sharper – almost 100% correct after just three examples – than might be expected with children who have not yet learned the relevant words in any languge. Clearly, it is very important to test these models on children directly, which is what I am now doing in collaboration with Fei Xu. Hopefully, applying the same formal model to both children and adult learners will help us to see what they have in common – what are the basic core abilities of human word learning – and also to make sense of the differences in computational terms.

In particular, the Bayesian modeling framework may help to answer important questions about the nature of the constraints and biases that children bring to the task of word learning. For example, in order to assess the contribution of a taxonomic bias, we could contrast the predictions of a Bayesian model using a taxonomic hypothesis space with those of Bayesian models using other kinds of hypothesis spaces, *e.g.* regions in a multidimensional space (as in Chapter 3) or overlapping clusters (as in Chapter 5), which could both be obtained by scaling similarity judgments. To take another example, Callanan et al. (1994) have argued that the basic-level bias suggested by Markman (1989) and others, and frequently observed in word learning experiments with children, is primarily a translation effect that occurs robustly only when children are shown examples for which they already know a basic-level name but no other name. In the present study, we found that adding a basic-level bias to the model's prior probability was helpful in accounting for participants' behavior, although this was not the only, or even the primary, source of explanatory power. Applying the Bayesian model to studies with children, using some objects with familiar basic-level names and others without, should help to resolve the translation issue, by establishing the degree to which a basic-level bias in the model's prior probability is necessary to account for children's generalization behavior in these different cases.

## 4.5 Rules vs. similarity *across* domains

> Why is a single instance, in some cases, sufficient for a complete induction, while, in others, myriads of concurring instances, without a single exception known or presumed, go such a very little way toward establishing a universal proposition? Whoever can answer this question knows more of the philosophy of logic than the wisest of the ancients and has solved the problem of induction.
>
> John Stuart Mill, *A System of Logic*

We have now seen two case studies of human concept learning in quite different domains, the abstract world of healthy levels and the realistic microworld of peppers,

trucks, and dogs. We saw that a basic phenomenon of learning emerged in both cases: convergence of generalization from a broad gradient (of similarity) in the earliest stages of learning to all-or-none generalization (by the minimal consistent rule) as more examples were observed. We also saw that a Bayesian framework – in particular, the Strong Bayes model – can explain how the learner makes this transition after seeing a sufficient number of positive examples of a concept.

In a much broader context, the convergence from uncertainty to certainty – the learning curve – is one of the classic themes of both human concept learning research (Shepard, Hovland & Jenkins, 1964; Bower & Trabasso, 1964; Nosofsky, Gluck, Palmeri, McKinley & Glauthier, 1994) and the Bayesian learning literature (Watanabe, 1960; Duda & Hart, 1973). A central question when studying learning curves has always been, "How fast?" What determines the rate of the learner's convergence to the target concept? Looking at our two case studies, we can see that the learning curves are actually quite different in each case. As Figure 19 illustrates, generalization gradients in the healthy levels task took between 10 and 50 examples to converge fully to the minimal rule (Chapter 3, Figures 7 and 14), while in the word learning task, generalization was practically all-or-none after only three examples!

We can also frame this task difference in terms of the rules vs. similarity debate. In the healthy levels task, the default mode of generalization from limited positive evidence appears to be graded, or similarity-based. By "default mode", I mean the manner of generalization after a few (*i.e.* three, not 1 or 100) examples have been observed. Given a few samples of healthy levels, people have a sense that other healthy levels must be similar to those, but they have no clear idea of where to draw a boundary. Each successive example causes the generalization gradient to sharpen up incrementally, converging after many examples to what looks like a rule, but no single rule ever pops out (either in the mind or in the data) as the clear criterion for generalization. There is no "aha" phenomenon typically associated with rule-based learning. In contrast, on the word learning task the default mode of generalization does appear to have an "aha", or rule-based, character. After just a few examples, we feel fairly confident that we know the actual criterion for using this new word, and

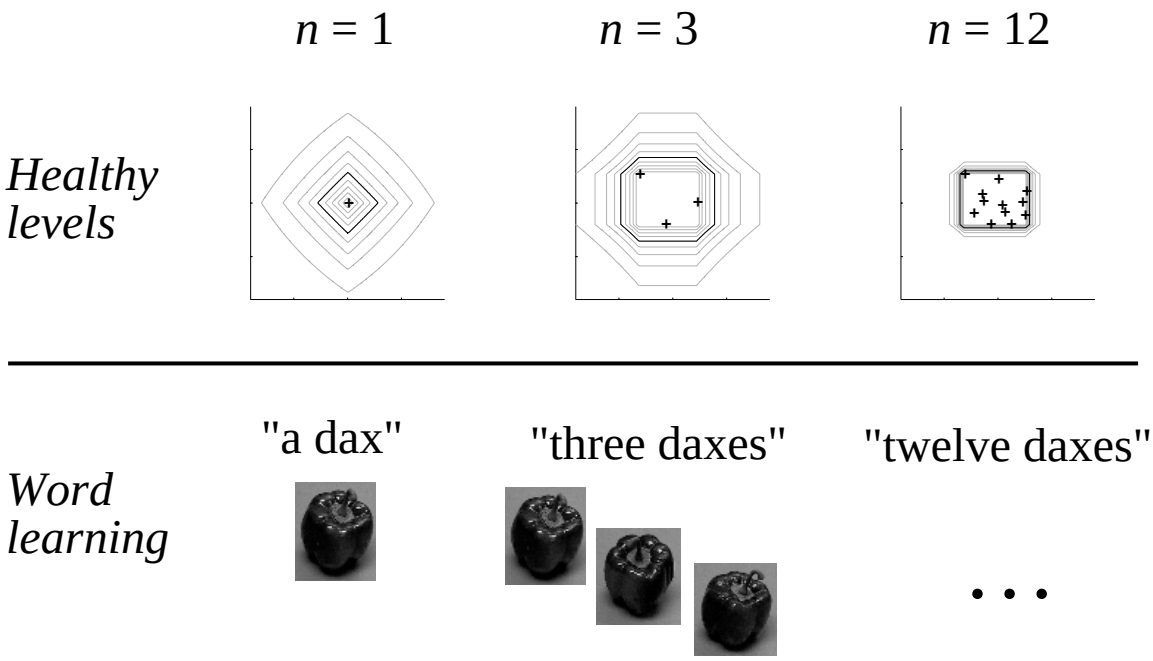



**Figure 19**

this is reflected in the very sharp generalization data collected. Introspectively, our uncertainty about how to generalize given just one labeled example feels more like the choice between several possible rules than like a primitive sense of similarity. Until now in the thesis, we have focused on understanding when and why rule-based versus similarity-based generalization emerges within each of these domains, but we can also ask the analogous question *across* domains: why do similarity gradients dominate in the healthy levels task and rules dominate in the word learning task?

It is possible that this question has no clean theoretical answer. The different patterns of learning observed in these two tasks may be merely due to accidental domain-dependent differences in how the mind/brain is wired up. Although at an abstract level, both the healthy levels and word learning tasks can be cast as problems of concept learning from limited positive evidence, that may be a mistaken generalization as far as the architecture of the mind is concerned. In other domains, particularly linguistic ones, there is good evidence for more than one fundamental learning mechanism (Pinker, 1991). Perhaps the mind is just set up to use rapid, rule-based mechanisms for learning words and slower, similarity-based mechanisms

for learning in continuous feature spaces, and that's the end of the story.

There are a few reasons not to give up so soon on a unified understanding of these phenomena. Learning to use words, unlike the acquisition of syntactic or morphophonological rules, probably relies on domain-general learning mechanisms (Markson & Bloom, 1997; Bloom & Markson, 1998); learning about the healthy levels of a substance almost certainly does. Also, we can switch the default modes in each domain by providing different kinds of input. In the healthy levels task, giving negative examples, *i.e.* unhealthy levels, can make a rule for discriminating positives from negatives seem perfectly natural. [9] In the word learning task, giving the same label to two objects $x_1$ and $x_2$ which appear to have nothing in common might lead to a last resort strategy of generalization based on similarity, *e.g.* "$x_1, x_2$ or anything very similar to one of those". And even if it is just a fact that one task is carried out in one brain module and the other task is carried out in a different brain module, we still would like to know why these tasks might be assigned to different modules, and why these modules work the way they do.

One intuitive answer is that the healthy levels task gives learners a continuum of possible ways to generalize, while the word learning task gives learners only a discrete set of choices. Perhaps similarity is the natural mode of generalization in continuous domains, while rules are the natural mode in discrete domains. However, this can't be true if taken literally. Recall the alternative version of the healthy levels task which used a one-dimensional numerical stimulus representation – *e.g.* "1400" and "1700" – instead of the graphical representation of dots in a two-dimensional space. In the numerical version, the set of possible generalizations is discrete and clearly enumerable: the minimum healthy level can be any number between 1000 and 2000, and the maximum can be any number between the minimum and 2000, for a total of $(1001 \times 1000)/2 = 500,500$ possible ways to generalize. Clearly, having a countable set of alternative generalizations is not the determinant of a rule-based default mode.

[9] Here are three random examples of levels that doctors consider healthy, 192, 195, and 198, and two random examples of levels considered unhealthy, 189 and 202. Exactly what the rule is that discriminates them is not certain, but the idea that there is a rule at all, *e.g.* positives lie between 190 and 200, is now intuitive.

It is also not the size (cardinality) of the set of possible generalizations that matters. In the word learning task, there are $2^{24} = 1,073,741,824$ possible ways to generalize, far more than in the numerical version of the healthy levels task.

Maybe it is the size of the learner's *hypothesis space* that makes the difference between the two tasks. In the healthy levels task, all of the 500,500 logically possible ranges of healthy levels seem like candidates for the hypothesis space, but in the word learning task, we used fewer than 40 clusters of objects as hypotheses. Intuitively, it seems that it must be easier to lock into a single hypothesis after just three examples when there are fewer hypotheses to choose from. This idea is on the right track, but is still not sufficient. Suppose that the concept to be learned corresponds to the width of a certain kind of machine part, and that you know that such parts are always manufactured with a width between two adjacent millimeter values (*e.g.* 452 and 453 mm), but you have no idea what those values are for this particular part. As in the healthy levels task, there is still a large number of possible hypotheses, but a single example of this kind of part (*e.g.* 738.2 mm) is now enough to lock in all-or-none generalization (738-739 mm). This case also eliminates another possibility, that the use of stimuli with transparent one- or two-dimensional content is somehow responsible for a default similarity-based mode.

One way in which both the word learning task and this part-size task differ from the original healthy levels task is in having no *partially overlapping* hypotheses. In the word learning task, the candidate extensions are nested – either disjoint or one wholly contained inside another – and in the part-size task they are completely disjoint, while in the healthy levels task each hypothesis overlaps many others to different degrees. Could this ambiguity be the source of the difference in how fast generalization converges to a rule? Not quite. Restricting the hypothesis space to a nested (rather than partially) overlapping structure is not on its own sufficient to induce rapid convergence to a rule. Recall the case of healthy levels of a poisonous substance (*e.g.* lead) that we considered at the end of Chapter 3. There, the hypothesis space was nested, in that each candidate extension for the range of healthy levels (from 0 up to some maximum level $m$) either contained or was contained in every other hypothesis.

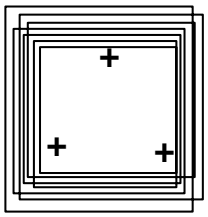
However, a gradient of generalization was still the norm after three examples.

The crucial feature that distinguishes tasks with fast convergence to a rule from those with much slower convergence is not the difference between nested and partially overlapping hypothesis spaces, but between *densely* and *sparsely* overlapping hypothesis spaces (Figure 20). That is, in all versions of the healthy levels task, each hypothesis overlaps many other candidate extensions with very similar sizes and boundaries, while in the word learning task, the extent of hypothesis overlap is much sparser (and is nonexistent in the part-size task). To understand how the density of hypothesis overlap impacts generalization behavior, we return to the Bayesian framework. As we will see, the same principles which we have used to explain the transition between similarity-like and rule-like generalization *within* individual tasks – hypothesis averaging and the size principle – also explain the relative dominance of these two modes of generalization *across* different tasks, as a function of the sparse or densely overlapping structure of the hypothesis space.

In the healthy levels task, the hypothesis space consists of all rectangular regions in the two-dimensional feature space. For any one rectangle hypothesis, there are many others which overlap with it almost completely. Some are a bit bigger, others a bit smaller, others the same size but slightly offset in position, and so on (Figure 20; see also Chapter 3, Figure 2b). As a result, any set of positive examples will always be consistent with many distinct hypotheses that differ only infinitessimally in size and location. This is what we mean in calling the hypothesis space "densely overlapping." Under the size principle for scoring the likelihood of each hypothesis, a densely overlapping system of hypotheses implies that any set of examples can always be explained by many possible hypotheses that all have roughly equal probabilities of being the true extension of the concept. Now, the probability of generalizing to new stimuli is determined by averaging the predictions of all consistent hypotheses weighted by their probabilties. The results of averaging many overlapping rectangle hypotheses of approximately equal probability are the broad gradients of generalization observed in Chapter 3, Figure 3. As more examples are observed, the posterior probability becomes more sharply peaked and the generalization gradient consequently becomes

*Healthy levels:*

**densely overlapping hypotheses**



---

*Word learning:*

**sparsely overlapping hypotheses**

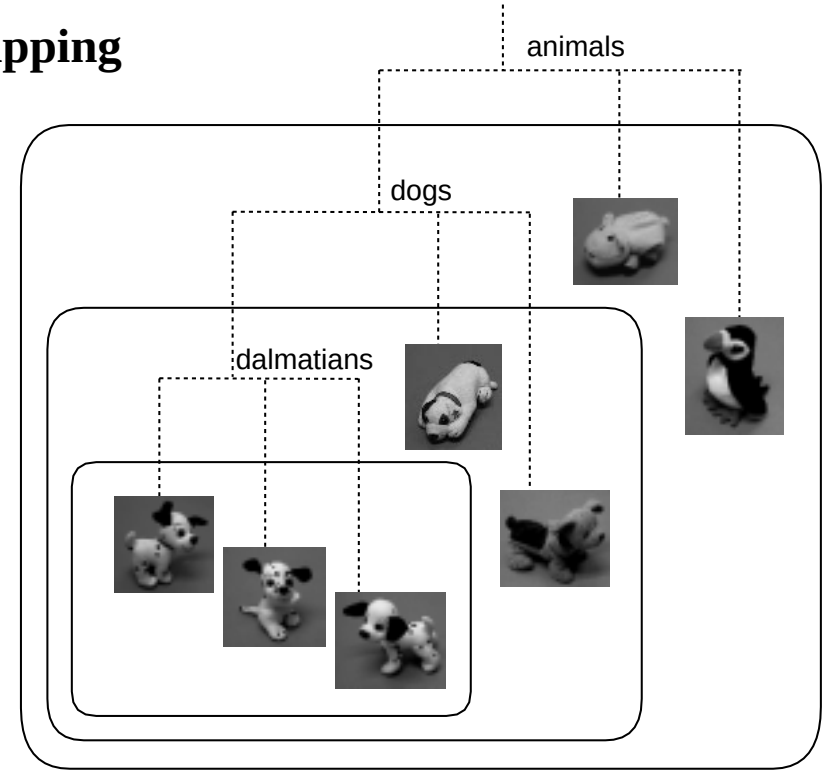



**Figure 20**

steeper. But because of the continuum of densely overlapping hypotheses, it takes many examples for the posterior to become so concentrated on the single smallest hypothesis that generalization appears to be all-or-none.

In the word learning task, the structure of the learner's hypothesis space is a very different one, which gives rise to qualitatively different patterns of generalization under the influence of hypothesis averaging and the size principle. Here, instead of a densely overlapping continuum of hypotheses, the learner has a much sparser hypothesis space in the form of a taxonomic hierarchy (Figure 20; see also Figures 6 and 7 of this chapter). It is still the case that any set of examples will be consistent with more than one hypothesis, but the number of consistent possibilities is much lower than in the rectangle hypothesis space. Moreover, each hypothesis is generally appreciably smaller than its parent, which is appreciably smaller than its parent, and so on. Thus the relative size differences between hypotheses, which turn into differences in likelihood via the size principle, are greater to start with and become compounded exponentially as we see just a few more examples. Observe that the greatest structural differences are found for hypotheses corresponding to the natural subordinate, basic, and superordinate classes in this domain, *e.g.* the set of all dogs (node #29), or the set of all yellow trucks (node #13), or the set of vegetables (node #9). These classes are very well separated from their parent classes in the tree, giving them not only a huge advantage in likelihood over their competitors, but also a high prior probability (which measures class distinctiveness, *i.e.* branch length). In sum, after seeing one example of a new word, all the classes above that object in the tree will be assigned some probability of being the word's true extension. Averaging over the predictions of these hypotheses gives rise to a graded pattern of generalization, as in Figure 5 (or Figure 17), row 1. But after just a few more examples are observed, the smallest consistent hypothesis becomes significantly more probable than any other hypothesis in the taxonomy, and generalization approaches an all-or-none function on that minimal rule.

Our explanation of the different default modes of generalization in these two tasks thus comes down to a difference in the structure of the relevant hypothesis spaces.

In the word learning task, sparse tree-structured hypotheses ensure that only a few examples will generally suffice to make one hypothesis significantly more probable than the others, leading to a rule-like pattern of generalization. In the healthy levels task, densely overlapping hypotheses ensure that even after a handful of examples, there will be many consistent hypotheses of roughly equal probability, and a gradient of similarity-like generalization will result.

The results of Feldman (1997) provide further support for this line of analysis. Feldman studied how people acquire simple perceptual categories in which the hypothesis space is even sparser than a taxonomic tree. These categories consisted of configurations of a line-segment and a dot, or two line-segments, in a two-dimensional display. Feldman argued that people's hypotheses in this domain formed a *lattice* (like a tree, but with multiple parents for each node) with two important properites: (1) each candidate extension was *infinitely* smaller than (*i.e.* measure-zero in) all larger extensions containing it; (2) there was always a unique smallest hypothesis containing any set of examples. Under the size principle, these two assumptions imply that the entire posterior probability should always be concentrated exclusively on the smallest consistent hypothesis, and hence that generalization should always be rule-based, even from a single example! That is more or less what Feldman (1997) found. Given just a single example of a dot on a line-segment, or two line-segments meeting at their endpoints, people restricted their generalizations to the most specific categories in over 90% of the trials.

It might be argued that this Bayesian analysis doesn't really *explain* the difference between these tasks so much as *locate* it in the structural features of the learner's hypothesis space, *i.e.* his prior knowledge. The question of where the learner's hypothesis space comes from is a much cloudier issue, which is hard to address without a complete theory of knowledge. It is certainly beyond the scope of the Bayesian framework for concept learning, or any other framework for concept learning, for that matter. (I will come back to this point in the final chapter's discussion.) What the Bayesian analysis *does* explain is how structural differences in prior knowledge – as opposed to differences in content, *i.e.*what the knowledge is about – lead to significant

observable differences in concept learning and generalization behavior. Locating the cause of the phenomenon in the form rather than the content of our knowledge is an important contribution, because, while there are an infinite number of different things we can think about, there are probably only a finite (and rather small) number of ways in which we can think about them. Understanding this link also allows us to predict the shape of generalization in some domain, given a hypothesis about how people's prior knowledge in that domain is structured, or to infer from people's generalization behavior the possible forms of their prior knowledge.

The third and final case study of this thesis will further explore the link between the form of prior knowledge and the shape of generalization, in a more complex domain where *multiple* kinds of prior knowledge are expected to give rise to multiple kinds of generalization behavior *within* a single domain. This domain is the number concept task, which served as our introduction to the challenges of concept learning in Chapters 1 and 2, and to which we now return in Chapter 5.

# Chapter 5

# Case Study #3: The Number Game

## 5.1 Introduction

This chapter's case study is the most speculative and exploratory in nature. I focus on a domain in which people have far more complex prior knowledge than we could hope to model with complete rigor, and explore the extent to which the Bayesian framework may nonetheless provide qualitative insights into the character of concept learning and generalization. The particular task is the number concept game introduced in Chapter 1. The learner's challenge here is to guess how a simple computer program will behave – specifically, which numbers between 1 and 100 it will accept – given one or more random examples of numbers that the program does accept. This domain may not be as "natural" a domain for concept learning as the microworld of objects in Chapter 4, but it is almost as rich – maybe more so – and far more amenable to analysis. Historically, some of the most insightful work on human and machine concept learning has focused on similar inductive inference tasks with numbers (Wason, 1960; Hofstadter, 1995), at least in part for similar reasons.

Recall that when we first analyzed this task in Chapter 2, we assumed a hypothesis space of only 30 possible concepts: odd numbers, even numbers, square numbers, all numbers (between 1 and 100), multiples of $j$ for $3 \leq j \leq 10$, numbers ending

in the digit $j$ for $1 \leq j \leq 9$, and powers of $j$ for $2 \leq j \leq 10$. We made this assumption purely for simplicity and convenience; the point there was to illustrate the Bayesian framework for concept learning, not to understand how people actually behave on this task. If we want to model real human behavior, the first step is to get some idea of what a realistic hypothesis space might look like. In Section 2, I approach this problem via an analysis of similarity in number domains, and outline the ingredients and predictions of a Bayesian concept learning framework in this vastly more complex hypothesis space. The analysis identifies two important subspaces of hypotheses for number concepts: one with a sparse structure like the taxonomic tree of Chapter 4, and the other with a densely overlapping structure like the rectangles of Chapter 3. This suggests a natural experiment to test my claim at the end of Chapter 4 about how the structure of prior knowledge influences the shape of generalization. If, as I argued, sparse hypothesis spaces lead to a default mode of rule-based generalization while densely overlapping hypothesis spaces lead to a default of similarity-based generalization, then we should be able to observe both kinds of generalization behavior within the single domain of number concepts. Section 3 presents an experiment designed to test this prediction. Section 4 considers how the Bayesian model might actually be implemented in the mind/brain using a combination of rule- and similarity-based heuristics, to avoid explicit computations over a >5000-element hypothesis space.

## 5.2 A Bayesian model of number concept learning

### 5.2.1 Hypothesis space

In principle, there are $2^{100}$ logically possible concepts in this domain. While the great majority of these subsets are probably not *psychologically* possible concepts, people's formal and informal knowledge about numbers is vast and multifaceted. It is hard to imagine even where to start looking for a hypothesis space of candidate extensions for human number concepts.

In the last chapter, we saw that a model of similarity judgment data was able to expose a reasonable hypothesis space for word learning. Perhaps the same approach might also allow us to uncover a natural hypothesis space for number concepts. In the word learning domain, the choice of a hierarchical clustering model for similarity was guided by important domain knowledge: the taxonomic constraint, and the idea that words (at least for natural object kinds) name clusters of similar things. In the number domain, it is less clear what the right model for similarity is. Some aspects of numbers seem taxonomic: *all numbers* contains *even numbers* and *odd numbers*; *even numbers* contains *multiples of four*, which contains *multiples of twelve*, and so on. However, a tree-structured hypothesis space is clearly not appropriate in general, because numbers frequently belong to two partially overlapping classes; *e.g.* 36 belongs to the square numbers and the even numbers. Other aspects of numbers seem to fit a spatial model of similarity, in particular along the important dimension of numerical magnitude. While that dimension would be naturally captured in a multidimensional scaling (MDS) model, discrete hypotheses like *square numbers* or *prime numbers* seem out of place in an MDS framework. We need a model that is more flexible than either hierarchical clustering or MDS, suitable for capturing the diversity of people's knowledge about number.

Shepard & Arabie (1979) introduced the *additive clustering* model of similarity for precisely these purposes, and convincingly demonstrated its application on a small data set of number stimuli collected by Shepard, Kilpatrick & Cunningham (1975). In earlier work (Tenenbaum, 1995), I translated the additive clustering model into a probabilistic framework and obtained somewhat better results on the number data set; I will review my results here. The additive clustering model is a version of Tversky's (1977) feature-based model of similarity (and thus a cousin of the Bayesian framework of this thesis – see Chapter 2), in which similarity is modeled as a weighted sum of the features common to two objects. Mathematically, we identify features with clusters, so that "having a feature" is equivalent to "belonging to the cluster of things with that feature". We assume that each stimulus $i$ is characterized by its membership in one or more of a set of $K$ clusters. We set the variable $f_{ik}$ equal to 1 if object $i$

| Rank | Weight | Stimuli in class | Interpretation |
|---|---|---|---|
| 1 | .444 | 2 4 8 | powers of two |
| 2 | .345 | 0 1 2 | small numbers |
| 3 | .331 | 3 6 9 | multiples of three |
| 4 | .291 | 6 7 8 9 | large numbers |
| 5 | .255 | 2 3 4 5 6 | middle numbers |
| 6 | .216 | 1 3 5 7 9 | odd numbers |
| 7 | .214 | 1 2 3 4 | smallish numbers |
| 8 | .172 | 4 5 6 7 8 | largish numbers |

Variance accounted for = 90.9% with 8 clusters (additive constant = .148).

**Table 1**

belongs to cluster $k$ (*i.e.* has feature $k$), and to 0 otherwise. Each cluster also receives a weight $w_k$, representing how salient or important it is in the similarity computation. We can then model the similarity $s_{ij}$ of two objects $i$ and $j$ as

$$s_{ij} = \sum_{k=1}^{K} w_k f_{ik} f_{jk}. \tag{5.1}$$

Because $f_{ik} f_{jk}$ equals 1 if and only if both objects belong to cluster $k$ (and otherwise equals 0), $s_{ij}$ is just the sum of the weights $w_k$ of features common to $i$ and $j$. Additive clustering algorithms search for a set of (real-valued) $w_k$ and (binary-valued) $f_{ik}$ parameters which make the predicted similarities from Equation 5.1 as close as possible (in squared error) to the observed similarity judgment data from an empirical study. The number of clusters $K$ is a free parameter.

Shepard, Kilpatrick & Cunningham (1975) asked subjects to judge the similarities of the "abstract concepts" of the numbers 0-9. Table 1 shows the results of submitting their data to the probabilistic version of additive clustering in Tenenbaum (1995). The output of the algorithm is just the weight and members of each cluster; the linguistic labels were added afterwards for convenient reference. As Shepard &

Arabie (1979) first pointed out, essentially two kinds of clusters occur in models of these data: those which seem to capture mathematical properties of numbers, such as *power of two* ($\{2, 4, 8\}$) or *multiple of three* ($\{3, 6, 9\}$), and those which refer to their numerical magnitude, such as *large numbers* ($\{6, 7, 8, 9\}$) or *smallish numbers* ($\{1, 2, 3, 4\}$). We can make several other observations as well: the clusters based on numerical magnitude always correspond to a connected interval of numbers, *i.e.* all numbers between some minimum and some maximum; the mathematical properties tend to have higher weight than the numerical magnitude features; and for both kinds of properties, the weight of a cluster seems to be inversely related to the number of stimuli it contains.

This division of number features into two families is supported by the developmental study of Miller & Gelman (1983). Miller & Gelman gave similarity judgment tasks to subjects in four different age groups, ranging from kindergartners to adults. They analyzed their data using INDCLUS, a variant of additive clustering in which subjects of all age groups were assumed to use the same features but the weights of the features were allowed to vary between age groups. As with the Shepard et al. (1975) data, clusters were found that corresponded to both mathematical properties and numerical magnitude properties. Moreover, the weights on these clusters varied as a function of subject age in one of two systematic ways, depending on whether the cluster expressed a mathematical property or a magnitude property. Specifically, the youngest children focused on magnitude properties almost exclusively. As subject age increased, the weights placed on magnitude properties diminished while the weights placed on mathematical properties increased, to the point where the mathematical properties tended to be weighted higher than magnitude properties by adult subjects.

How do these results help us in constructing a hypothesis space for number concept learning? A natural assumption is that each feature recovered by clustering the number similarity data corresponds to a candidate extension for the computer program in the number concept task. Note that this is the direct analog of how we constructed a hypothesis space in the word learning domain, only with a more expressive model of similarity that allows for partially overlapping clusters. Moreover,

I propose to identify the weight of each cluster with the prior probability of the corresponding hypothesis. Assigning likelihoods is much easier here in than in the word learning domain, because we have a clear and concrete measure of the size of each hypothesis, *i.e.* how many numbers it contains, and the likelihood given $n$ consistent examples is just $[1/\text{size}(h)]^n$. Thus we have all the ingredients needed for a Bayesian model of number concept learning: hypothesis space, priors, and likelihoods.

There's just one problem here, and it's a big problem. This analysis only gives us a hypothesis space for the numbers 0-9, and we want a hypothesis space for the numbers 1-100. Most of the relevant hypotheses over the domain 1-100 have only one – or zero! – instances in the domain 0-9. In a concrete sense, then, these additive clustering models are practically useless for the task we want to model. Directly generalizing this approach to cover the whole domain 1-100, *e.g.* by having subjects judge the similarities of all pairs of numbers less than 100 and submitting those data to additive clustering, would be impossible; the time requirements for both human subjects and the computer algorithm are prohibitive.

However, relaxing our standards of rigor somewhat, there is still a way to use these results to guide us in constructing a reasonable hypothesis space for the domain 1-100. In the above analysis, I identified several characteristics of the simple hypothesis space generated by additive clustering on the domain 0-9. We can try to specify, by hand, an analogous set of hypotheses over the domain 1-100 that displays these same characteristics. The relevant specifications are: one set of hypotheses captures the salient mathematical relationships between numbers; another set of hypotheses captures the dimension of numerical magnitude in the form of overlapping intervals of numbers; the mathematical properties tend to have higher weight than the numerical properties; smaller hypotheses tend to have higher weight. These characteristics of the hypothesis space for 0-9 are not just an arbitrary list or an accidental outcome of the Shepard et al. (1975) study, but were replicated in even stronger form in the Miller & Gelman (1983) result.

Table 2 shows the specific hypothesis space $\mathcal{H}$ used to construct a Bayesian model of learning on the number concept task. The particular choices made were obviously

**Hypothesis space for number game**

Mathematical properites:

- Odd numbers
- Even numbers
- Square numbers
- Cube numbers
- Primes
- Multiples of $n$: $3 \leq n \leq 12$
- Powers of $n$: $2 \leq n \leq 10$
- Numbers ending in $n$: $0 \leq n \leq 9$

Magnitude properties:

- Intervals between $n$ and $m$: $1 \leq n \leq 100$; $n \leq m \leq 100$

**Table 2**

arbitrary, but they follow the general patterns observed in the additive clustering studies. As suggested by those studies, the hypotheses in $\mathcal{H}$ fall into two groups. Mathematical properties are represented by hypotheses for odd numbers, even numbers, square numbers, cube numbers, and prime numbers, along with multiples and powers of small numbers, and numbers ending in the same digit. Numerical magnitude is represented, as it was in the domain 0-9, by hypotheses corresponding to all intervals of numbers contained within the domain 1-100. The total number of hypotheses is 5090.

### 5.2.2 Priors and likelihoods

As we have done throughout the thesis, we take the likelihood of hypothesis $h$ given $n$ examples $X = \{x_1 \ldots x_n\}$ to be

$$p(X|h) = 1/|h|^n,$$

or 0 if $h$ does not include all of $X$. This is just the usual strong sampling generative model, embodying the assumption that the examples are a random sample from the concept's true extension, and gives rise to the size principle for scoring hypotheses. The basic intuition for the size principle in the number concept domain was developed at length in Chapter 2, so I do not recap it here.

The Bayesian model also requires a prior probability distribution over the hypothesis space. Rather than trying to assign prior probabilities to each of 5090 hypotheses individually, a hierarchical approach was adopted. First, the hypotheses were divided into two groups corresponding to mathematical and magnitude properites. A certain fraction $\lambda < 1$ of the probability was allocated to the mathematical properties as a group, leaving $1 - \lambda$ for the magnitude properties. Within the group of mathematical properties, that $\lambda$ probability mass was distributed uniformly (*i.e.* giving a constant probability to each hypothesis in this group). Within the group of intervals representing numerical magnitude, the $1 - \lambda$ probability mass was distributed as a function of interval size, just as for the rectangle hypotheses in Chapter 3. The specific size prior used was an Erlang distribution with expected size parameter $\sigma$ (Equation 3.10). The Erlang distribution was chosen to capture the intuition that the computer program is more likely to pick out an interval of some intermediate size (determined by $\sigma$), rather than of very small or very large size. The values of $\sigma$ and $\lambda$ are treated as free parameters of the model.

The idea of allocating prior probability at a top level to the two groups of hypotheses and only then subdividing it within groups was suggested in part by the intution that this was how subjects would cognize this task, and also by the way the weights of features in the Miller & Gelman (1983) developmental study covaried as if

they belonged to these two independent groups. This way of assigning prior probability also instantiates an automatic bias in favor of the mathematical rules over the magnitude-based hypotheses, because there are many more magnitude hypotheses than mathematical hypotheses. That is, there are only 40 mathematical hypotheses, so each one receives a prior probability of $\lambda/40$, but there are 5050 interval hypotheses, which each receive an average of $(1-\lambda)/5050$ in prior probability, or roughly $1/100$ of what the mathematical hypotheses each receive assuming $\lambda$ is near 0.5. This preference can be seen as a version of the size principle, now applied to the prior instead of the likelihood. The entire model can be seen as a hierarchical three-stage generative process for the examples: in the first stage, a choice is made between mathematical rules or magnitude intervals as the basis for the concept; in the second stage, a specific rule or interval is chosen as the concept's extension; in the third stage, a specific set of examples is chosen from the extension. The parameters $\lambda$ and $\sigma$ enter at different stages of this process, stage 1 and stage 2 respectively, both of which are considered to be part of the "prior". The size principle operates at all three stages, giving a preference to outcomes which are sampled from smaller sets over outcomes which are sampled from larger sets. Appendix C considers several other hierarchical generative models for concept learning and their impact on generalization behavior in a Bayesian framework. Constructing hierarchical hypothesis spaces such as these one may be a general strategy for making the Bayesian framework applicable to richly structured real-world domains.

Finally, we compute the probability of generalization to a new object $y$ – the probability that the program will accept $y$, given the random examples $X$ of numbers it does accept – by averaging the predictions of all our hypotheses about the program's extension, weighted by their posterior probabilities $p(h|X)$:

$$p(y \in C|X) = \sum_{h \in \mathcal{H}} p(y \in C|h)p(h|X). \tag{5.2}$$

Figure 1 shows this model's generalization behavior for three different sets of examples: $\{16\}$, $\{16, 8, 2, 64\}$, and $\{16, 23, 19, 20\}$. (The values of the two free pa-

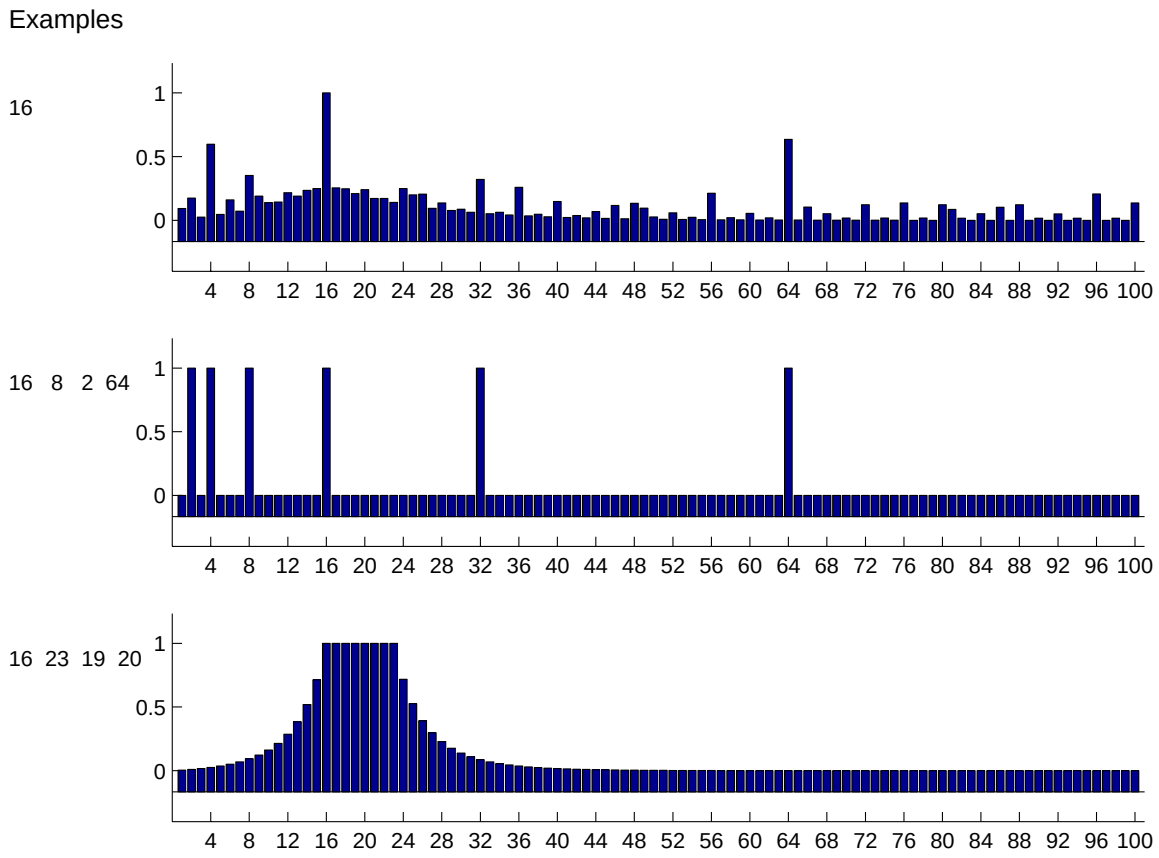



**Figure 1**

rameters $\lambda = 2/3$ and $\sigma = 10$ were set by eye.) The first two example sets, and the Bayesian learner's behavior, are familiar from Chapters 1 and 2. Generalization from 16 alone is very fuzzy and uncertain; after 8, 2, and 64 are also observed, generalization converges to an all-or-none function on the most specific rule, *powers of two*. The behavior on the third set is quite different from either of these. Given 16, 23, 19, and 20, the model now predicts a broad gradient of generalization to numbers of similar magnitude, rather than convergence to any one rule. Figures 2, 3, and 4 show how these patterns emerge from averaging over all consistent hypotheses, as a function of how peaked the posterior probability is.

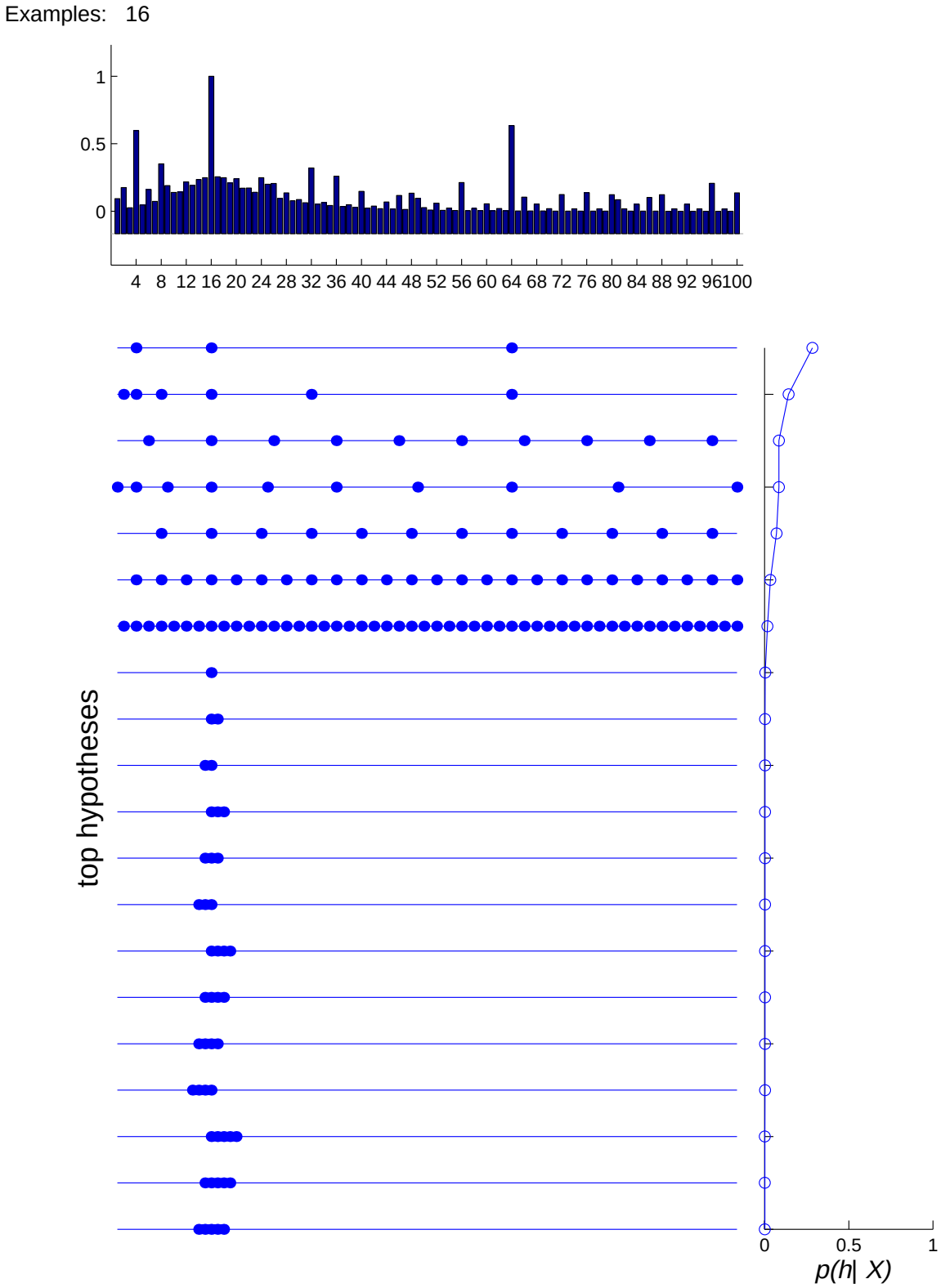

Examples: 16
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96100
top hypotheses
0
0.5
1
p(h| X)

**Figure 2**

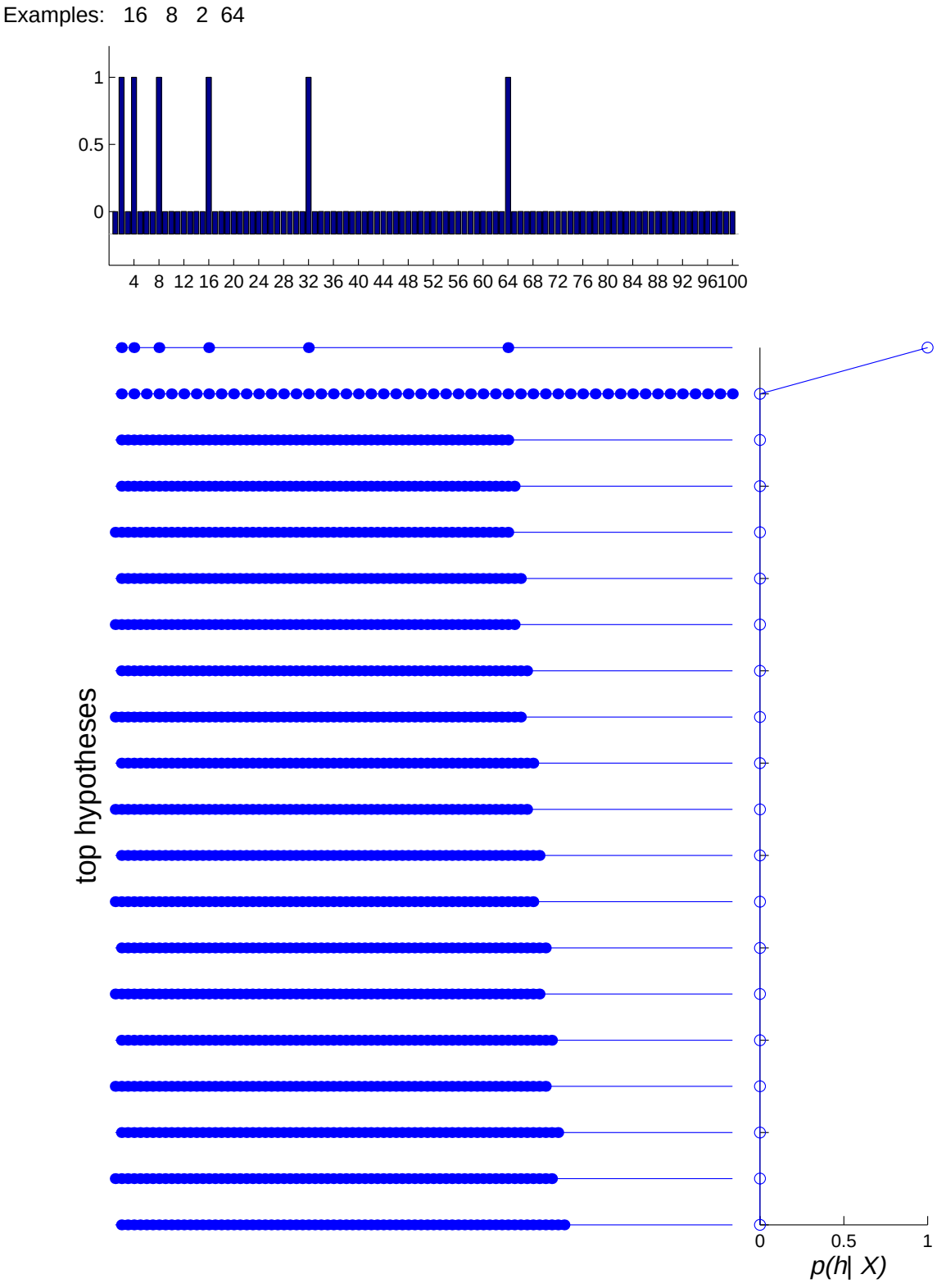

Examples: 16 8 2 64
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96100
top hypotheses
0
0.5
1
p(h| X)

**Figure 3**

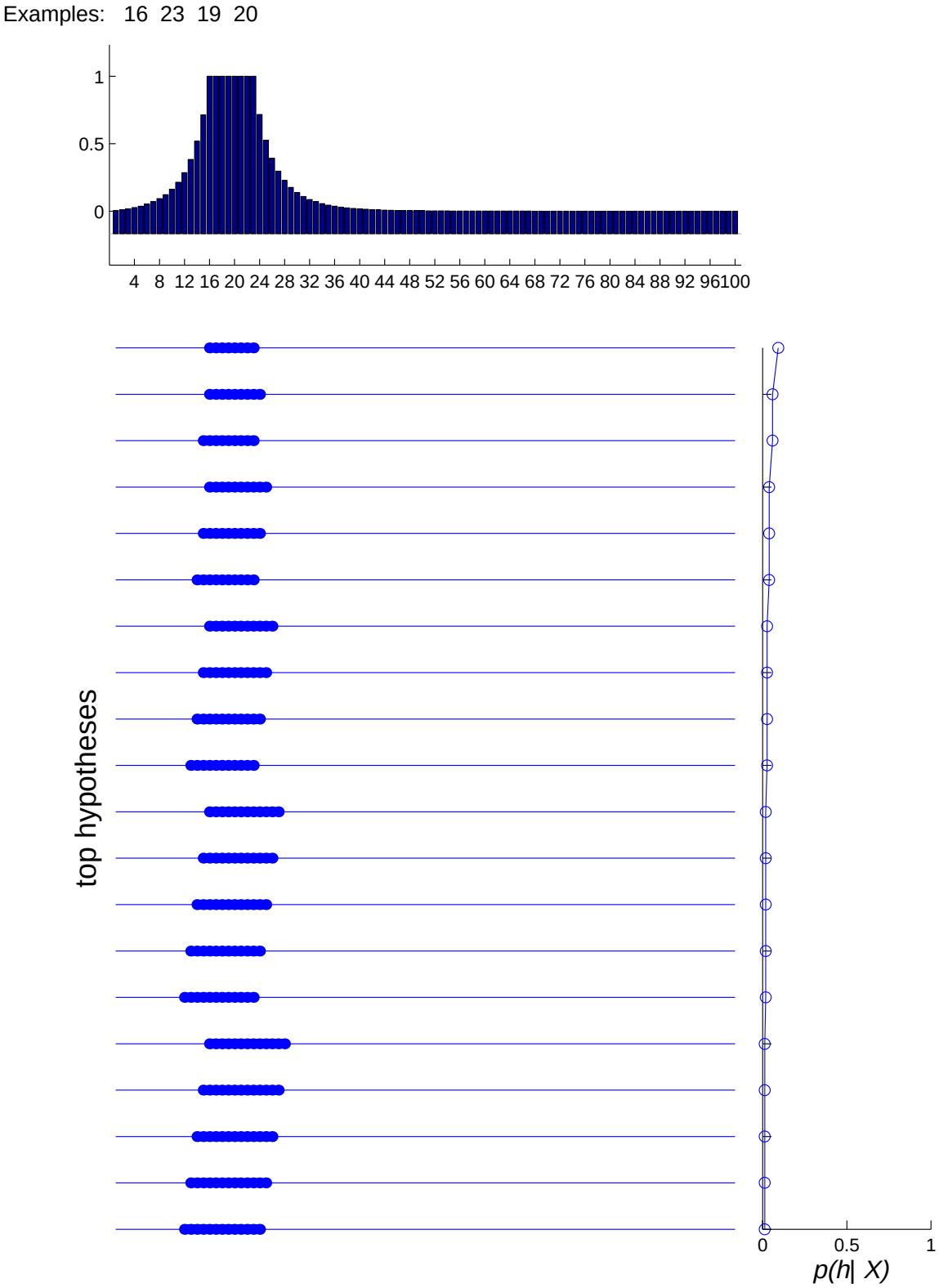

Examples: 16 23 19 20
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96100
top hypotheses
0
0.5
1
p(h| X)

**Figure 4**

### 5.2.3 Patterns of generalization: rules *and* similarity

In each version of the Bayesian framework until now, the transition from initial similarity-like generalization to subsequent rule-like generalization always seemed to follow a single, stereotyped course. In the word learning task, three examples were typically sufficient to induce all-or-none generalization; in contrast, in the healthy levels task, generalization was still quite graded after only three examples, and somewhere between 10 to 50 examples were required before those gradients tightened up to the minimal consistent rule. This model, for the first time, contains both modes of generalization as potential defaults. Which one is activated depends on the particular examples observed. Given a few examples consistent with one or more known mathematical rules, the model locks into the most specific rule with all-or-none behavior (Figure 1, row 2). Otherwise, the model generates a gradient of generalization based on similarity in numerical magnitude to the examples (row 3), which sharpens up only gradually as successive examples are observed. Given only one example, the model has committed to neither of these modes; in expression of its ambivalence, it assigns roughly uniform and rather low generalization probabilities to many numbers (row 1).

The source of this model's more flexible behavior can be traced to the more complex structure of its hypothesis space. At the end of Chapter 4, I attributed the emergence of rule-based or similarity-based default modes of generalization in Bayesian inference to the distinction between sparse and densely overlapping hypothesis spaces. In a sparse hypothesis space (as in the word learning task or Feldman's (1997) perceptual categorization task), the smallest consistent hypothesis is appreciably smaller than all other consistent hypotheses; under the influence of the size principle, it rapidly becomes the only probable way to generalize after just a few examples have been seen. In contrast, with a continuum of densely overlapping hypotheses (as in the healthy levels task), the smallest consistent hypothesis is surrounded by many others of quite similar size and hence similar likelihood; only after a relatively large number of examples have been observed does the posterior probability become sufficiently

concentrated to produce rule-like generalization.

The case of number concepts is more complex because people's a priori candidate extensions include *both* sparsely structured mathematical classes – *even numbers*, *multiples of four*, *powers of two*, etc. – as well as densely overlapping intervals based on numerical magnitude – *numbers between 15 and 25*, *numbers between 15 and 30*, *numbers betwen 14 and 28*, and so on. Each of these hypothesis space structures is capable of giving rise to its stereotypic generalization pattern – rules or similarity – depending upon whether it is engaged by the observed examples. Moreover, the sparse mathematical hypotheses will tend to take precedence over the dense interval hypotheses, because of how the they include numbers which are quite far apart in magnitude. For instance, the set of examples 16, 8, 2, and 64 is consistent with both mathematical hypotheses (*even numbers*, *powers of two*) and interval hypotheses (*numbers between 1 and 65*, *numbers between 1 and 80*, etc.), but the smallest mathematical hypothesis, *powers of two*, contains only 6 numbers (between 1 and 100), which is much smaller than the smallest interval hypothesis, *numbers between 1 and 65*. Under the influence of the size principle, this leads to a difference in probability that is compounded exponentially with each new example, with the result of rapid convergence to the rule *powers of two*.

The precise predictions of the Bayesian model will obviously vary depending upon exactly which hypotheses are included and how the prior probability assignment is made. However, the qualitative pattern illustrated in Figure 1 – that either rule-like or similarity-like generalization may emerge after several examples have been seen – is always predicted, as long as the learner's hypothesis space has the dual structure uncovered in the additive clustering studies, of sparse mathematical hypotheses combined with densely overlapping intervals representing numerical magnitude. The experiment in the following section was designed to test this qualitative prediction with human learners.

## 5.3 An experimental test

### 5.3.1 Methods

Eight people participated in the study. Participants were members of the broad MIT community. All gave informed consent and were compensated for their participation.

Participants were given a description of the number concept task very much like that at the beginning of Chapter 1. Just as in Chapter 1, they were first given a few examples of the kinds of concepts that the computer might be programmed to implement: "X is even", "X is between 30 and 45", "X is a power of 3", "X is less than 10". To minimize the possibility of bias, only these four sample concepts were mentioned and they were split two and two between mathematical properties and magnitude properties.

The example sets were designed to test the Bayesian model's prediction that either rapid convergence to a rule or a slowly sharpening gradient of similarity can emerge after a few examples are observed, depending on the nature of the examples. Specifically, we designed three pairs of four-example sets, with the two sets in each pair based on the same "seed" number but designed to evoke either rule-based or similarity-based generalization. One pair was the same as in Figure 1, based on the seed 16: $\{16, 8, 2, 64\}$ and $\{16, 23, 19, 20\}$. The other two were based on 60, $\{60, 80, 10, 30\}$ and $\{60, 52, 57, 55\}$, and 81, $\{81, 25, 4, 36\}$ and $\{81, 98, 86, 93\}$ To these 6 trials, we added two trials with just a single example each, 16 or 60. These eight trials occured in one of four pseudo-random orders, subject to the constraint that the two single-example trials always occured before the six four-example trials.

On each trial, participants were shown one of these sets of examples as "random examples of numbers that program A [or B or C or ...] accepts". They were then given a list of 30 probe numbers between 1 and 100, and asked to rate each one on a scale of 1 to 7 according to how probable it was to be accepted by the program. Subjects were instructed to use the scale consistently across trials to represent their degree of confidence. To encourage consistent use of the scale, it was reprinted on each page of the experimental booklet along with suggested probability-of-acceptance

interpretations for each level: 1, "less than 5%"; 2, "10%"; 3, "30%"; 4, "50%"; 5, "70%"; 6, "90%; 7, "greater than 95%". The particular choice of 30 probe numbers was somewhat subjective, but was guided by the model predictions to try to include all the numbers with a reasonably high predicted probability of acceptance, along with several numbers with low probabilities to anchor the scale. The same 30 probe numbers were used on each trial with the same seed example, *i.e.* trials using the example sets $\{16\}$, $\{16, 8, 2, 64\}$, and $\{16, 23, 19, 20\}$ all used one set of probes, trials using the example sets $\{60\}$, $\{60, 80, 10, 30\}$ and $\{60, 52, 57, 55\}$ used a second set of probes, and so on. Each trial took up one page of an experimental booklet, and probe numbers appeared in a random order on that page. Participants were told they could rate the probe numbers in any order they wanted to.

At the conclusion of the experiment, participants were asked to describe in words the set of numbers which they thought each program accepts.

### 5.3.2 Results

Figure 5 shows the average data from the eight participants. Individual participants' data, in most cases, looks similar but noisier. The height of each bar represents people's average judgments of how likely the corresponding number is to be accepted by the computer program, given the one or four random examples listed on the left of the plot. The data were scaled linearly from the 1-7 rating scale to the interval $[0, 1]$, for comparison with the predicted probabilities. Note that there are only 30 bars on each plot, representing the 30 probe numbers that participants judged. Thus a missing bar does *not* mean a judgment of zero probability; it means that no judgment was collected. (A few of the data bars are also "made up", because not all numbers that were included in a trial's example set were also included in the probe set for that trial. However, people always gave ratings of 7 to probe numbers that were also examples (with the exception of only a single judgment on one trial), so it seems reasonable to extrapolate a mean judgment of 7 for the few probe numbers that were all also examples but were not actually rated by subjects.)

Figure 6 shows the predictions of the Bayesian model for the same example sets.

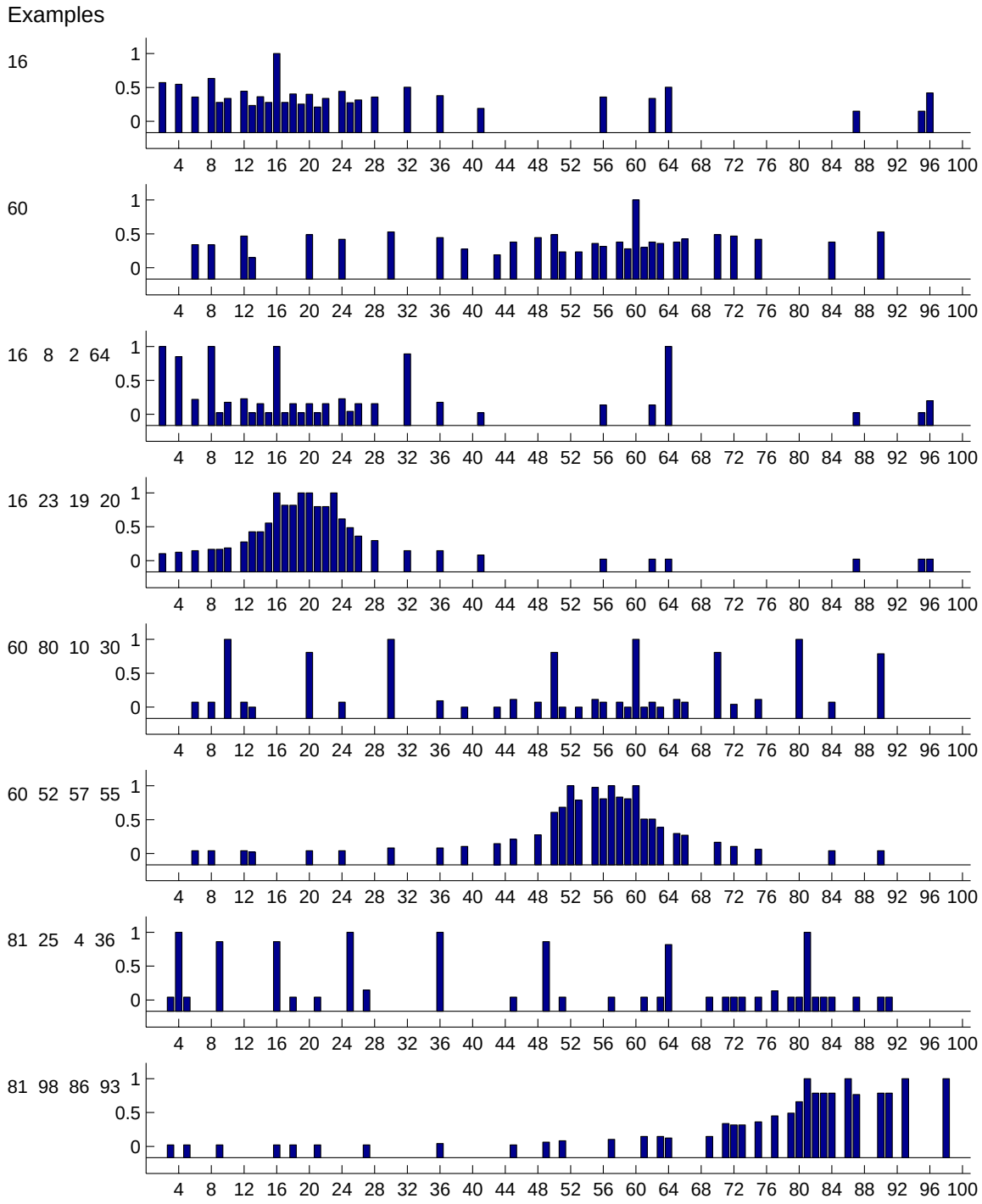

Examples
16
60
16 8 2 64
16 23 19 20
60 80 10 30
60 52 57 55
81 25 4 36
81 98 86 93
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100

**Figure 5**

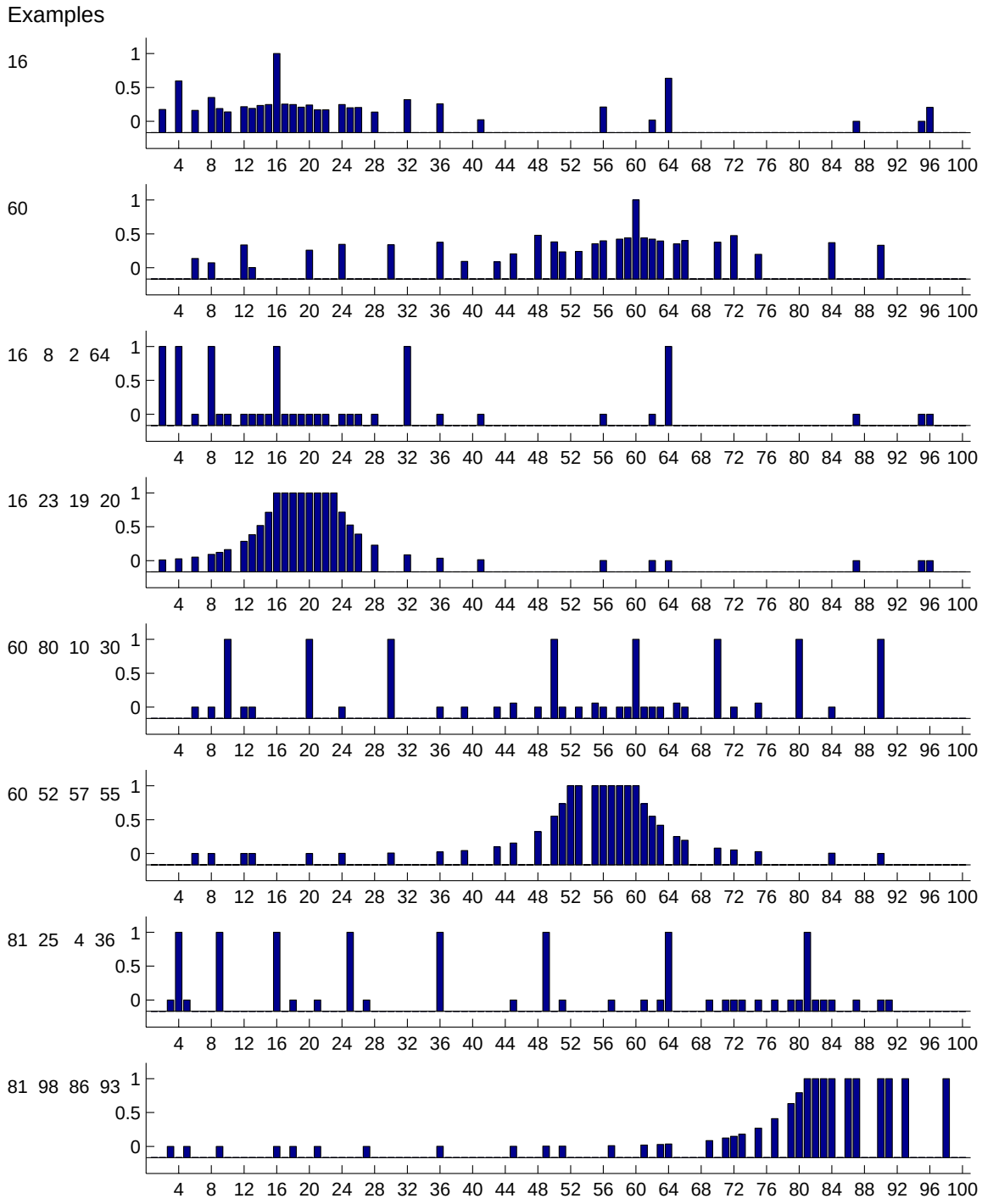

Examples
16
60
16 8 2 64
16 23 19 20
60 80 10 30
60 52 57 55
81 25 4 36
81 98 86 93
1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100

**Figure 6**

The model parameters $\sigma$ and $\lambda$, the same as those used in Figure 1, were chosen based on intuition before the experiment was conducted. (These values could almost certainly be adjusted to improve the a posteriori fit of the model, although I have not done that yet.)

### 5.3.3 Discussion

The results are in qualitative agreement with the model predictions and, in some cases, quite close quantitative agreement as well. Given only one example (Figure 6, rows 1,2), people gave the probe numbers fairly uniform probabilities of acceptance, typically lower than 0.5 but significantly greater than zero. The model did likewise, although with a lower baseline acceptance rate. Given four examples, people's generalization judgments followed one of two basic patterns just as predicted. For example sets consistent with one or more simple mathematical properties, ($\{16, 8, 2, 64\}$, $\{60, 80, 10, 30\}$, and $\{81, 25, 4, 36\}$, people gave essentially all-or-none generalizations based on one of these properties (rows 3, 5, and 7). For example sets not consistent with any simple mathematical property, $\{16, 23, 19, 20\}$, $\{60, 52, 57, 55\}$, and $\{81, 98, 86, 93\}$, people generalized according to a gradient of similarity based on numerical magnitude (rows 4, 6, and 8).

More specific predictions were borne out as well. In two of the three rule-based cases, the examples were consistent with *more than one* simple mathematical rule (*i.e.* $\{16, 8, 2, 64\}$ is consistent with "even numbers" as well as "powers of two", $\{60, 80, 10, 30\}$ is consistent with "multiples of five" and "even numbers" as well as "multiples of ten"), but people generalized almost exclusively in accordance with the *most specific* rule, as predicted by the theory. On trials with broad gradients of generlization, the theory predicted montonically decaying gradients of generalization with approximately the correct rate of decay. Like the prediction that rule-based generalization converges to the minimal rule, the predicted rates of decay of these gradients are a consequence of the size principle (as opposed to the value of $\sigma$ in the prior on interval size, which counts for rather little after four examples are observed).

The several ways that model and data differ are also quite interesting. In particu-

lar, the model assigns 100% probability-of-acceptance to any probe number consistent with all of the same hypotheses that the observed examples are consistent with. In contrast, people always assigned a rating of 7 to probe numbers which were also given as examples on the same trial, but very rarely assigned a 7 to any other probe number. One way to interpret this behavior is that people are just using the difference between 6 and 7 to show the difference between very high and perfect certainty. Another interpretation is that people are actually allowing for other possible explanations of the examples which may be less natural mathematically, but still logically possible. This second interpretation is quite sensible; after all, there are $2^{96}$ subsets consistent with any four examples, and there is a reasonable expectation that the experimenter might be trying to trick them! It could also be incorporated into the Bayesian model, *e.g.* by allowing for a small portion of the prior probability to be distributed over all $2^{100}$ logically possible hypotheses.

Another interesting point of difference shows up in two trials with broad gradients of generalization, $\{60, 52, 57, 55\}$ (row 6) and $\{81, 98, 86, 93\}$ (row 8). People's judgments, rather than decaying with a smoothly changing, always-concave slope as the model predicts, show several inhomogeneities, where the slope of the gradient appears to change suddenly. These sudden changes occur at suspicious numbers: 60 and 50 in row 6, and 70 in row 8. It appears that people did not consider all numerical intervals equally, but gave special attention to those with minimum and/or maximum values at a multiple of ten. Looking at how participants described their judgments verbally confirms this interpretation. It does seem quite intuitive that the multiples of ten could serve as the "cognitive reference points" (Rosch, 1975) for judging similarity of magnitude in the domain 1-100, and thus cause categorical distortions in the underlying similarity metric. As retailers have long known, \$3.99 is much more similar to \$3.00 than is \$4.01. This sort of effect could be incorporated into the Bayesian model by placing significantly higher prior probability on interval hypotheses that are bounded by multiples of 10.

## 5.4 Heuristics for Bayesian concept learning

The key finding of the experiment is conclusive. Within the context of a single concept learning task and a single domain of stimuli, people will adopt either rule-like or similarity-like strategies for generalization depending on the particular examples that they observe. I presented a model that explains these two modes of behavior – and why they occur when they do – as special cases of a single underlying computation: Bayesian inference over a hypothesis space of candidate extensions of the concept. As in our previous case studies, the size principle was the major force driving the dynamics of both rule-like and similarity-like generalization, as well as the transition between them. The new feature of this case study was the presence of both sparse and densely overlapping subspaces of hypotheses, which gave rise, under the influence of the size principle and hypothesis averaging, to the possibility of either rules or similarity as default modes of generalization.

Participants in our experiment were not prompted to use rule-based, similarity-based, Bayesian, or any other strategies for generalization. They were told nothing about hypothesis spaces with thousands of candidate extensions, sparse versus densely overlapping hypotheses, hypothesis averaging, or the size principle. The only instructions they were given were that the computer accepts some numbers and not others, that they should judge the probability of new numbers being acceptable under the assumption that the given examples were random samples from the acceptable set, and that the computer was programmed to pick out sets like *even numbers*, *numbers between 30 and 45*, *powers of 3*, or *numbers less than 10*. Everything else came from inside their heads, and came out more or less in accord with the predictions of the rational Bayesian model.

Despite this success, it is not at all clear that the actual processing going on inside people's minds during this task looks very much like the specifics of the Bayesian model. At the level of conscious processing, it is almost certainly impossible for people to keep thousands of candidate hypotheses in mind and update the probability of each one in accordance with Bayes' theorem, as each new example is observed. At the level

of unconscious processing, it is much harder to say what people can and cannot do, yet it still seems unlikely that people are explicitly evaluating the relative probability of two hypotheses as different as *all powers of two* and *all numbers between 13 and 29*. When debriefed after the experiment, people typically reported following a much more heuristic approach. Given one example, they generalized to probe numbers on the basis of how many mathematical properties they had in common with the example, and, to a lesser degree, their similarity in magnitude. Given four examples, people went with a single mathematical rule if it "popped out" at them, otherwise they went with a graded sense of similarity based on numerical magnitude.

In light of this, the Bayesian framework will only be complete as a psychological theory if it makes contact with – and ideally, justifies – the heuristics people naturally use to learn concepts from examples. In the rest of this section, I will show how the three major components of people's heuristic strategy on this task each embody a justifiable approximation the full Bayesian model, in the situations where people apply them. More deeply, the knowledge that people have about *when* to apply these different heuristics – at least as important as the individual heuristics themselves – is itself rationally justified by the Bayesian framework.

### 5.4.1 The MIN heuristic and the hazaka principle

First consider people's rule-based behavior. After seeing a few examples consistent with a simple rule, people chose that rule as the basis for all-or-none generalization. When the examples were consistent with multiple rules, as in two of the three cases here, people overwhelmingly went with the most specific one. This can be seen as instance of the MIN algorithm for learning rule-based concepts from only positive examples. I reviewed the history of this algorithm in Chapter 2, so let me just recall the two most important points here. First, MIN is not only a proposal for human concept learning, but is also one of the classic algorithms for inductive inference and reasoning more generally. Accounts of language acquisition (Pinker, 1995; Wexler & Manzini, 1987), scientific inference (Popper, 1959), and folk inference (Osherson et al., 1990), as well as theories of machine learning (Valiant, 1984; Haussler, 1988),

have all drawn on MIN as a crucial ingredient. Second, the classical justifications of MIN as a rational inference procedure are only valid *asymptotically*: *i.e.* in the limit of infinite (or sufficiently many) data, it allows the learner to converge (or get arbitrarily close) to the true concept from only positive evidence.

An intriguing finding of this experiment is that people are intuitively prone to using the MIN algorithm given just a few examples. After four examples of the program's acceptable numbers, $\{16, 8, 2, 64\}$, people generalized only to the other powers of two, even though the perfectly natural class of all even numbers was still consistent with these observations. We found essentially the same result in Chapter 4; given three green peppers as examples of "blickets", people generalized only to other green peppers and no further. In both tasks, people seem to consistently apply an inference procedure that is justified only for asymptotically large samples when they have seen just three or four data points. From the classical point of view, it looks like we have uncovered a new version of the "law of small numbers" (Tversky and Kahneman, 1971) – the mistaken belief that the law of large numbers applies equally well to samples with small numbers of data points.

The Bayesian framework of this thesis, however, shows that the law of small numbers actually applies in many concept learning tasks, including these! Under the assumptions that 16, 8, 2, and 64 are all randomly sampled from the concept, and that one of our hypotheses is in fact the true extension, generalizing strictly according to the minimal rule *powers of two* is the rational way to behave. Similarly, under the assumptions that three green peppers are a random sample from the extension of the word "blicket", and that "blicket" maps onto one of the candidate extensions in our hypothesis space, then restricting the meaning of "blicket" to just green peppers is the rational thing to do. Of course, either of these applications of MIN could be called irrational if the learner has reason to doubt his hypothesis space, or to doubt that the examples are in fact random samples. But whether those assumptions are justified is a separate issue from whether the learner's inference algorithm is justified.

A crucial aspect of any good heuristic is the knowledge of when to apply it. The Bayesian framework justifies the application of MIN for some cases of small

samples, but by no means all. Remarkably, people seem to know intuitively when MIN is justified and when it is not. In general, neither people nor Bayes think MIN is valid given just a single example, otherwise we would see all-or-none rule-based generalization from just 16, or just a single green pepper. But people do realize – correctly! – when the observation of just three or four examples consistent with a restrictive rule provides strong support for the rule's general applicability.

Many bits of folk wisdom embody the same realization, with a particular focus on the number three. When we say, "the third time's the charm!", we mean that if something is not working, it's worth trying it three times before giving up on it for good. The Talmud considers it reasonable to assume that a practice which has been carried out in the same way three times will always be carried out that way. The three consistent examples are said to establish a *hazaka*, or "propensity", that is legally binding. States which passed "three strikes and you're out" laws in the 1990's, putting away people for life after three felony convictions, were convinced that three examples were enough to establish someone as a habitual felon. The "Hazaka Principle" and its variants seem to be a ubiquotous means of generalization in many everyday reasoning situations, not just in concept learning.

All of these "hazaka"-type principles can be seen as having the same underlying probabilistic basis as the MIN heuristic in concept learning. Any one outcome of a process has many possible explanations, with some being simpler, *i.e.* more probable, than others. Let's say that the simplest explanation of an outcome is that such an outcome will always occur, and that this explanation has a higher probability (likelihood) than any other explanation by a factor of $K$. After three identical outcomes, more than one explanation is still possible, but if the events are independent, then the simplest explanation – that the observed outcome always occurs – becomes much more likely than the alternatives. Mathematically, this is because the probabilities of independent events *multiply*; the simplest hypothesis is now $K^3$ times more likely than its nearest competitor. If on one example, the simpler explanation is twice as likely as the next best alternative to be true, then it is eight times more likely after three. If it was originally three times more likely, then it is now 27 times more likely!

There are also the exceptions that prove the rule. These are cases where Bayes justifies faster or slower acceptance of the minimal rule than after three examples, and people intuitively follow the same principles. In the healthy levels task of Chapter 3, both the Bayesian model and the participants in our study knew that the minimal consistent rectangle was not a good generalization after only three or four examples. On the contrary, 10 to 50 examples were needed for convergence to the minimal rule. In the language of the previous paragraph, the densely overlapping hypothesis space means that $K$ – the likelihood advantage of the smallest consistent rectangle – is only infinitessimally greater than 1. Thus it takes many examples for $K^n$ to become significantly greater than 1. In Feldman's (1997) perceptual categorization experiments (discussed at the end of Chapter 4), we have the opposite situation: K is infinite, because the smallest consistent hypothesis is infinitely smaller than (technically, measure zero in) any others. Bayes then predicts that just a single example is enough to validate the minimal consistent rule, and people agreed in over 90% of the trials. Together, these cases are quite reminiscent of Nisbett and colleague's response to the "law of small numbers" phenomenon (Nisbett, Krantz, Jepson & Kunda, 1983). They argued that people in fact had an appreciation for the effect of sample size on the validity of universal generalizations, by showing that people's intuitions about necessary sample sizes varied in systematic and normatively sensible ways across different domains of knowledge.

In sum, people routinely generalize by accepting the minimal rule consistent with their observations. But they also have strong intuitions about how many observations are necessary before accepting the minimal rule. The required number of observations varies from context to context as a function of the structure – sparse or densely overlapping – of the hypothesis space of candidate explanations. And at least in every case investigated in this thesis, people's judgments of when to apply the MIN heuristic have been more or less in line with the rational standards of Bayesian inference.

### 5.4.2 Similarity heuristics

Generalizing in an all-or-none fashion via the most specific consistent rule is only one component of people's strategy on the number concept task. The two other components can both be described as similarity heuristics. Given just one example, people generalize in a graded fashion based on how much a probe number overlaps with the example on a combination of mathematical and magnitude features. Given four examples not consistent with a simple mathematical rule, people generalize based on how similar a probe number is to the examples along the dimension of magnitude only. These two similarity heuristics correspond to the two classic ways of thinking about and modeling similarity in the literature, as a contrast of common and distinctive features (Tversky, 1977) or as distance in a single or multidimensional space (Shepard, 1980). Throughout the thesis, I have tried to make the point that both featural and spatial notions of similarity can be seen as special cases of the Bayesian framework for generalization. But now we are starting to see why this unification is significant. We found in this study that both featural and spatial intuitions of similarity may occur to people within the context of a single task. The Bayesian analysis predicts when – and explains in rational terms *why* – a featural or spatial model of similarity should be most appropriate to describe people's generalization behavior.

Viewing both senses of similarity in Bayesian terms also gives us insight into the flexibility and dynamics of similarity computations. Tversky (1977) first focused researchers' attention on these questions, by providing a model of similarity that could accomodate many of the subtle ways that similarity can vary from context to context. However, Tversky's model does not *predict* these flexibilities of similarity, it merely *accomodates* them with the proper settings of certain free parameters. The topic of flexible similarity is the center of much of contemporary research as well (Medin, Goldstone & Gentner, 1993; Goldstone, 1994, Medin & Florian, 1995), but it is still the case that formal models (*e.g.* Nosofsky, 1986; Ashby, 1992) focus on accomodating the phenomena of flexible similarity rather than predict them. Models of classification learning (Kruschke, 1992; Aha & Goldstone, 1992) are a notable exception, but as I

pointed out in Chapter 4 and in Appendix A, these models require both positive and negative examples of a concept to learn. The Bayesian framework, for the first time, gives us a way to understand and predict some aspects of flexible similarity to one or more positive examples of a concept.

In the case of spatial similarity, which (inspired by Shepard, 1987) we model using densely overlapping regions in a continuous space, the Bayesian framework predicts how gradients of similarity will shrink or stretch depending on the number and distribution of the observed examples. We were able to understand these deformations of the generalization gradients in terms of a rational inference process, rather than (or really, as a complement to) the conventional explanation in terms of "increasing stimulus discriminability" and "selective attention". This topic was discussed extensively in Chapter 3, so I will say no more about it here.

In the case of featural similarity, we now have an objective explanatory account of several phenomena which were only accomodated in Tversky's (1977) model by adjusting free parameters. [1] Recall the form of Tversky's *contrast* model:

$$\mathrm{SIM}(y \rightarrow x) = f(X \cap Y) - \alpha f(X - Y) - \beta f(Y - X). \tag{5.3}$$

Here objects $x$ and $y$ are represented by sets of binary features. (In the context of number concepts, these features might be things like *multiple of two*, *less than 10*, and so on.) $X \cap Y$ denotes the *common* features of both $x$ and $y$; $X - Y$ denotes the distinctive features of $x$ not shared by $y$. A *measure* $f(S)$ is assigned to each feature set $S$; generally $f$ is an additive measure, so that each feature $h_k$ in $S$ receives some weight $w_k$ and then $f(S) = \sum_{h_k \in S} w_k$. Thus the similarity of $y$ to $x$ is given by a contrast between their common and distinctive features, weighted by the parameters $\alpha$ and $\beta$. To accomodate the fact that the salience of features changes depending on the stimuli under comparison, Tversky allows $f(S)$ to be a free parameter (perhaps indirectly via the $w_k$ terms). To accomodate the fact that similarity may be asymmetric, Tversky

[1] The Bayesian model also has free parameters. The difference is that these parameters are assumed to be set once for each task and not to vary from judgment to judgment depending on the particular examples observed.

allows $\alpha$ and $\beta$ to be different and to change depending on the judgment context. But the model has no formal account of how these parameters are set.

The Bayesian framework for concept learning is not a model of similarity per se, but it does capture, in a more rigorous form, some of these phenomena as they occur when generalizing concepts based on similarity. Recall the final form of the Bayesian generalization function derived in Chapter 2 (Equation 2.5):

$$p(y \in C|X) = \frac{\sum_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n}{\sum_{h \in \mathcal{H}_X} p(h)/|h|^n}. \tag{5.4}$$

Here $\mathcal{H}_X$ denotes the set of hypotheses consistent with all the examples in $X$, and $\mathcal{H}_{X,y}$ denotes the set of hypotheses consistent with $X$ and also the new object $y$. Another way to write this is

$$p(y \in C|X) = \frac{1}{1 + \frac{\sum_{h \in \mathcal{H}_{X,\bar{y}}} p(h)/|h|^n}{\sum_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n}}, \tag{5.5}$$

where $\mathcal{H}_{X,\bar{y}}$ denotes the set of hypotheses consistent with $X$ *but not consistent* with $y$. If we identify hypotheses with features, then Equation 5.5 has an intuitive interpretation very similar to Tversky's model.[2] Each feature $h$ is assigned a weight that is a product of two terms: $p(h)$ and $1/|h|^n$. The expression $\sum_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n$ is just the sum of the weights of all features common to the examples $X$ and the new object $y$. The expression $\sum_{h \in \mathcal{H}_{X,\bar{y}}} p(h)/|h|^n$ is the sum of the weights of all features common to the examples but *not* shared with the new object, *i.e.* the distinctive features of the examples. Finally, the ratio of these two sums in the denominator means that similarity will increase with the weighted sum of features common to both $X$ and $y$ and will decrease with the distinctive features of $X$.

The most obvious difference between Equation 5.5 and the contrast model is that the contrast model is only defined for similarity to a single example, whereas the Bayesian model works for any number of examples.[3] But even for just a single example

[2]In fact, Equation 5.5 also satisfies all of Tversky's qualitative axioms for a *matching function*, the most general form of his model (Tversky, 1977).

[3]Heit (1997) extends the contrast model to multiple examples on non-Bayesian heuristic grounds

$X = x$, the Bayesian formulation make strong predictions where Tversky's model has only free parameters. Where the contrast model assigns each feature $h$ a single flexible weight $w_k$, we now have a product of the two terms $p(h)$ and $1/|h|^n$. The first term measures the a priori naturalness of a feature, and this is just as unconstrained as $w_k$ in the contrast model. But the second term implies that all other things being equal, features with smaller extensions will have larger weight. Tversky (1977) discusses this phenomenon – which he calls the *extension effect* – but gives no formal explanation for its occurence. The second term also implies that features with smaller extensions will become increasingly more salient as more examples are observed, as in the increasing importance of *power of two* relative to *even number* for guiding generalization after 8, 2, and 64 have been added to the one example 16. The contrast model has nothing to say about this.

The Bayesian framework also predicts an asymmetry of generalization without having to change any parameters like $\alpha$ or $\beta$. If one object $x$ has more distinctive features than a second object $y$, then Equation 5.5 predicts that generalization from $x$ to $y$ will be lower than generalization from $y$ to $x$. To test this prediction, we gave subjects in the number concept experiment a short post-test which asked them for direct judgments of the asymmetry, inspired by Tversky's (1977) famous question to subjects, "Which would you prefer to say, 'Red China is similar to North Korea' or 'North Korea is similar to Red China'?" In our case, subjects were given a random number $x$ that program A accepts and a random number $y$ that program B accepts. They were then asked whether they thought it more likely that program A also accepts $y$ or that program B also accepts $x$, and to rate the strength of their preference on a scale of 1 to 4. The Bayesian model predicts strong asymmetries for pairs of numbers $x$ and $y$ when $x$ has distinctive mathematical features that $y$ lacks, *e.g.* 24 vs. 26, or 81 vs. 89. Each subject saw 12 trials, on 6 of which the Bayesian model predicted an asymmetry and on on 6 of which it predicted no asymmetry (*e.g.* 74 vs. 76, or

and arrives at a similar formulation to this one. He also notes a possible connection to Bayesian inference, although he does not pursue the implications we discuss here. Heit (1998), however, does give a Bayesian rationale for asymmetry phenomena like the one discussed below.

53 vs. 57). [4] The results overwhelmingly confirmed the predictions, both in terms of which trials produced an asymmetry and the direction in which the asymmetry went. Participants showed a significantly stronger preference for one direction of generalization on those trials for which an asymmetry was predicted than on those trials for which no asymmetry was predicted, and they also showed the predicted direction of asymmetry on all six of the asymmetric trials. More details about this experiment will appear in a forthcoming paper.

I do not mean to imply that all, or even most, phenomena of flexible and dynamic similarity can be captured in this Bayesian model. This is a model of learning and generalization, not similarity, and it only bears on people's intuitions of similarity to the extent that they figure into concept learning. Outside of that task, the rational analysis that motivated the model does not necessarily apply. But within the context of learning concepts from one or more examples, the Bayesian model makes strong predictions about the context dependence of similarity that previous models have been able only to accomodate.

### 5.4.3 Putting rules and similarity together

For all the insight that a Bayesian framework gives us into individual heuristics based on rules or similarity, the most important thing it explains is how those heuristics work together to guide concept learning and generalization from very few positive examples, and why they work the way they do. Previous models of concept learning have postulated hybrid combinations of rule- and similarity-based components on intuitive grounds (Nosofsky & Palmeri, 1998; Erickson & Kruschke, 1998; Osherson et al., 1990) or neuroanatomical considerations (Ashby et al., 1998). Each of these models assumes different mechanisms for rule and similarity modules, and different

[4] The 6 trials on which an asymmetry was predicted were: 24 vs. 26, 36 vs. 34, 25 vs. 29, 72 vs. 68, 64 vs. 58, and 81 vs. 89. In each case, the first stimulus was more mathematically distinctive than the second stimulus, and thus generalization was predicted to be greater from the second to the first stimulus than from the first to the second stimulus. The 6 trials on which no asymmetry was predicted were: 74 vs. 76, 93 vs. 91, 53 vs. 57, 82 vs. 78, 52 vs. 46, and 11 vs. 19. In the actual experiment, the orders of the trials and of the two stimuli within each trial were pseudo-randomly varied.

means by which they interact. From a computational point of view, all of these proposals are somewhat ad hoc and it is difficult to see why an organism would be built along the lines of one proposal rather than another. The Bayesian framework of this thesis allows us to understand rules and similarity as complementary adaptations to a single computational challenge faced by human learners, to see when and why one mode or the other will dominate, and how each mode works to solve its part of the computational problem. Whether or not rule- and similarity-based mechanisms for concept learning are implemented in separate brain modules is a fascinating empirical question, for which evidence is just beginning to come in (Smith et al., 1998). But regardless of which way the brain story unfolds, we will only be able to make sense of it in the context of a unified computational framework for concept learning.

# Chapter 6

# Summary and conclusions

## 6.1 Overview of the Bayesian framework

I began this thesis by posing the fundamental computational problem of concept learning: how can people acquire such a rich range of possible concepts from the very limited evidence – one or a few positive examples – that they normally encounter? I then proposed a solution based on the principles of Bayesian inference. The Bayesian framework has four major ingredients, which we saw illustrated on three different kinds of concept learning tasks: continuous separable feature spaces (the "healthy levels" task), word learning, and the number game.

The first ingredient, in common with many classical approaches to induction, is to assume a *hypothesis space* of candidate extensions for the concept to be learned. Without some kind of restriction on the hypotheses we consider, generalization from any finite evidence – not to mention only a few positive examples! – is impossible. In a continuous separable feature space of stimuli, such as the healthy levels domain, a natural hypothesis space consisted of all rectangular regions in that space. In learning words for objects, the hypotheses corresponded to taxonomic classes in a hierarchy of object kinds. In learning number concepts, the hypotheses included both mathematically special classes, such as *all powers of two*, and sets of numbers with similar magnitudes, such as *all numbers between 10 and 20*. Each of these hypotheses can be thought of as a possible "micro-rule" for generalizing the concept, or alterna-

tively, as a candidate "feature" that could distinguish instances of the concept from noninstances.

The second step in the Bayesian framework is to assign a *prior probability* to each element of the hypothesis space. The prior embodies our beliefs about which hypotheses are the most likely candidates for new concepts in general, independent of any examples we have seen. In some sense, the hypothesis space itself is an extension of the prior; excluding logically possible hypotheses from our hypothesis space is equivalent to including them but assigning them a prior probability of zero. However, the prior allows us to encode finer degrees of preference. Over a hypothesis space of rectangular regions in feature space, our prior distribution might embody the knowledge that concepts tend to have a known typical size, and thus give preference to hypotheses with approximately that size. Over a hierarchy of taxonomic classes for word learning, our prior assumed that words map onto highly distinctive classes of objects (as defined by high within-class similarity), and also, perhaps, classes at the psychologically "basic" level. In the number game, our prior assigned higher weight to mathematically special sets of numbers, like *square numbers* or *even numbers*, relative to more psychologically generic interval-based hypotheses, like *numbers between 12 and 32*.

The third ingredient is a *generative model* of the examples, which allows us to score hypotheses based on their likelihood of having produced the data that we observed. Throughout, we made the assumption of *strong sampling*: the examples are a random sample from the concept's true extension. Strong sampling leads to the *size principle* for scoring hypotheses, which is how we were able to decide between two (or more) possible generalizations, each natural *a priori* and each consistent with the data. The size principle says that smaller hypotheses are more likely to be the true concept than larger hypotheses, and they become exponentially more likely as the number of consistent examples increases. We saw the dramatic effect of the size principle in each of the three case studies. Given any one example – a particular healthy level, a green pepper, the number 16 – there was little reason to prefer more specific hypotheses – a small rectangle, *green peppers*, *powers of two*, – over more general hypotheses – a large

rectangle, *peppers*, *even numbers*. But after we observe more examples consistent with both general and specific hypotheses – a tight cluster of healthy levels, three green peppers, the numbers 16, 8, 2, and 64 – the more specific hypothesis in each case appears to be the better explanation. The intuition is a statistical one: just as we would be much less likely to encounter a tight cluster of points if we were sampling from a large region than if were sampling from a smaller region, so would we be much less likely to encouter three green peppers sampling from *all peppers* as opposed to *all green peppers*, or 16, 8, 2, 64 sampling from *all even numbers* as opposed to *all powers of two*.

Finally, our actual generalization behavior in the Bayesian framework is determined by the *posterior probability* and the principle of *hypothesis averaging*. The posterior probability of each hypothesis is equal to the product of its prior probability and size-based likelihood. This gives the rational degree of belief in each hypothesis as a function of both our prior knowledge about more or less natural candidate extensions and the statistical information carried by the examples. Then, in order to decide the probability that any new object belongs to the concept, we average the predictions of all our hypotheses, weighted by their posterior probabilities. Intuitively, this means that we add up the weights (*i.e.* posterior probabilities) of all hypotheses consistent with both the examples and the new object, and compare this sum to the total weight of all hypotheses consistent with the examples regardless of whether or not they include the new object. The ratio of these two sums (which is always less than or equal to unity, because the second sum necessarily includes every term in the first sum although the converse usually does not hold) gives the probability of generalizing from the examples to the new object. Depending on how many different hypotheses receive significantly high posterior probability, our generalizations will be more or less graded. When the posterior is spread out broadly over many hypotheses, our generalization behavior will really be an average of all of these possible "rules", and thus will follow a gradient of similarity defined in terms of the number of hypotheses/rules/features that a new object shares with the examples. When the posterior is concentrated on a single hypothesis, on the other hand, the weighted average over

all hypotheses only pays attention to the extension of that one best hypothesis, and generalization becomes an all-or-none matter of consistency with this rule.

These are the four basic components of Bayesian concept learning, but to really understand the theory and its implications, it is necessary to see how these components interact as a function of the kinds of prior knowledge learners bring to their task and the kinds of data they observe. I explored these issues over the course of three case studies and reached several conclusions, recapped in the next section.

## 6.2 Summary of major contributions

### 6.2.1 How is concept learning even possible?

First and foremost, the Bayesian framework offers a solution to the problem of induction in concept learning that has eluded the classical approaches to generalization via abstract rules or similarity to exemplars. Both strict rule- and similarity-based approaches run into difficulty with the *flexibility* of generalization. Rule-based approaches need some way to explain how and why rules change their rankings in light of the observed examples; similarity-based approaches need to explain how and why the features or dimensions of similarity rise or fall in importance depending on the data. This was the focus of our discussion in Chapter 1. The Bayesian framework asserts that these are just two ways of looking at a single underlying problem with a single solution: statistical inference. To the extent that this is a solution to the problem of induction, it is a solution not in the sense that the inferences reached are guaranteed to be right – Hume's problem of induction – but in the sense that it offers a principled characterization of how people can learn the kinds of concepts that they actually do from the kinds of evidence that they actually observe – (one aspect of) Goodman's "new" problem of induction. The Bayesian framework thus takes us a step closer to the integration of the theory and practice of concept learning, which so many psychologists and computer scientists are working to achieve.

### 6.2.2 Quantitative modeling of generalization data in diverse domains

The case studies in Chapters 3-5 presented quantitatve data on how people generalize concepts from one or a few positive examples in three diverse task domains. For each task – drawing rectangles, learning words, guessing number concepts – we developed a model within the Bayesian framework that gave very accurate fits to the behavior of human learners. The major assumptions of the Bayesian model (the hypothesis space, the prior, and the measure of hypothesis size) were as independently motivated as possible, with a minimum of free parameters. To my knowledge, this is the first such attempt to model quantitatively the course of human concept learning and generalization from one or a few positive examples, using a single framework that applies across quite different task domains.

### 6.2.3 The appearance of rule-like or similarity-like generalization

A theme running throughout the three case studies was the relation between two kinds of generalization behavior, all-or-none, rule-like generalization and graded, similarity-like generalization. Both patterns occurred in each task domain, depending on the number (and sometimes the distribution) of examples observed. The Bayesian framework was capable of predicting when – and explaining why – generalization appears to follow a rule in some situations and a gradient of similarity in others. We looked at this question both *within* individual learning tasks and *across* different tasks.

Within particular learning tasks, we found the number of examples observed to be the crucial variable in determining the rule- or similarity-like character of generalization. In the earliest stages of learning – after we have seen just a few healthy levels, or a single labeled object in the case of word learning – the probability of generalization decayed gradually as a function of the appropriate measure of similarity to the observed example(s). In later stages – after 10 to 50 examples in the healthy levels task, but only 3 in the word learning task – generalization approached an all-or-none

function of the most specific rule consistent with the examples. The Bayesian framework explained this transition in terms of hypothesis averaging and the size principle, ingredients three and four above. Hypothesis averaging leads to sharp or fuzzy generalization behavior depending on whether the posterior probability distribution is peaked or flat; the size principle dictates that the posterior becomes concentrated on the single smallest consistent hypothesis, as the number of examples increases.

Not only within individual learning tasks, but across different task domains, we also found a difference in the relative dominance of rule- or similarity-based modes of generalization. Similarity-based generalization seemed to be the norm in the healthy levels tasks of Chapter 3, with convergence to the minimal rule occuring relatively slowly, after around 10 - 50 examples. In the word learning task of Chapter 4, in contrast, generalization was only graded given a single labeled example; as soon as only three examples were observed, rule-based generalization emerged clearly as the norm. The Bayesian framework explains this difference across domains in terms of structural features of the learner's hypothesis space. The space of rectangular regions in a continuous feature space – our hypothesis space in the healthy levels task – is a *densely overlapping* structure, with many hypotheses of similar sizes and boundaries consistent with any set of examples. In contrast, the hierarchy of taxonomic classes in the word learning task, particularly for those classes corresponding to natural subordinate, basic, or superordinate categories, is a *sparse* structure, with the most specific consistent hypothesis typically being signficantly smaller than any other consistent hypotheses. Because the likelihood of a hypothesis is a function of its size, dense hypothesis spaces will tend to give rise to broader posterior probability distributions, while sparse hypothesis spaces will lead to posterior probability being more concentrated on the single smallest hypothesis. Through the mechanism of hypothesis averaging, this in turn means that graded generalization will be more typical of dense hypothesis spaces while all-or-none, rule-based generalization will be more the norm in sparse spaces. Exactly in the same way that a small hypothesis becomes exponentially more probable than a large hypothesis as we see more examples, the relative advantage of the smallest hypothesis in a sparse space becomes exponentially

greater than the relative advantage of the smallest hypothesis in a dense space, as we see more examples. This explains why it can take as many as 50 examples to approach rule-like behavior in the dense space of rectangular regions in the plane, while just 3 examples were sufficient to achieve all-or-none generalization in word learning. Finally, we made the prediction that a more complex hypothesis space with both dense and sparse components should be able to give rise to either rule- or similarity-based generalization from the same number of examples, depending on the particular values of the examples observed. This prediction was borne out by our third case study, on the number game.

In sum, hypothesis averaging and the size principle together bring enormous explanatory power to the questions of when and why generalization of a concept appears governed by rules or similarity. Hypothesis averaging allows the Bayesian framework to contain both rules and similarity depending on the width of the posterior probability distribution; the size principle (modulated by the sparseness of the hypothesis space) determines when generalization will be based on rules or similarity by controlling the width of the posterior. Previous Bayesian approaches to concept learning have contained one or the other of these ingredients, but not both together. Putting them together has allowed us to see how rules and similarity complement each other as the two natural poles of rational concept learning.

### 6.2.4 Formulating rule-based and similarity-based heuristics in Bayesian terms

In parallel with showing that the Bayesian framework can explain both rule-like and similarity-like generalization behavior, we also showed how classic theories of concept learning based on rules or similarity can be viewed as special cases of Bayesian models, exactly valid in certain limiting cases and approximately valid as heuristics outside of those limits. By formulating strictly rule-based or similarity-based algorithms in terms of the Bayesian framework, we gained insight into their rational basis and their domain of justified validity, as well as how to make them more flexible or applicable

in new contexts.

The classic rule-based algorithm MIN – which generalizes according to the most specific rule consistent with the examples – is conventionally justified only asymptotically, in the limit of infinite data. We showed that given the right kind of hypothesis space – *i.e.* a sparse structure – and the proper generative model – *i.e.* strong sampling – MIN is justified as a good approximation to the Strong Bayes algorithm after only three or four examples have been observed. We found that people have an intuitive (if often unconscious) awareness of the importance of these requirements and a subtle sensitivity to whether or not they are satisfied in any given situation. When MIN does quickly become a good approximation to Strong Bayes, as in the word learning and number concept tasks, people do in fact stick to the minimal consistent hypothesis after three or four examples. In more extreme cases, such as when MIN is theoretically justified only after a larger number of examples (as in the healthy levels task), or, at the other extreme, after only a single example (as in Feldman's (1997) perceptual categorization experiments), people generalize accordingly. Finally, we noted that people's intuition that the most specific candidate generalization consistent with the data becomes valid after approximately three confirming instances are observed appears not only in concept learning, but also in many informal heuristics for social inference – such as the "hazaka" principle from the Talmud or the "three strikes and you're out" laws of many U. S. states.

The notion that similarity – if allowed to be flexible in the proper ways – could form the foundation of our concepts has been a popular position in psychology, but lacks substance without a rigorous model of how exactly the similarity computation is supposed to flex as a function of the observed examples. In each of the three case studies of this thesis, I presented a Bayesian model that could be interpreted as the principled mechanism behind a flexible similarity-based approach to concept learning. In the healthy levels task, the Bayesian model's predictions of how generalization gradients in feature space deform as a function of the examples observed provides a rational explanation for what was previously explained only in informal, process-oriented terms, such as "increasing stimulus distinctiveness" and "selective attention".

In the second case study, the Bayesian model represents a proposal for extending learning models based on flexible similarity to apply to domains with real objects as stimuli (represented in terms of a taxonomic hierarchy), a real task (learning the meaning of a word) as the goal, and the real demand of having to learn from only one or a few positive examples. In the number game, the Bayesian model explained how the relative weights of mathematical versus magnitude properties (or what we might think of as deep versus superifical causes of similarity between two numbers) varied as a function of the examples observed. It also predicted the presence and direction of asymmetries of generalization for pairs of numbers. Tversky's (1997) classic *contrast model* of flexible similarity can accomodate the phenomena of variable feature salience and generalization asymmetry, but does not predict them as the Bayesian model does.

Understanding how the Bayesian framework relates to traditional models based strictly on rules or similarity may also be relevant to mapping out the psychological and physiological processes responsible for concept learning. Our Bayesian models frequently employed very large hypothesis spaces, for which the necessary computations of scoring each alternative hypothesis and integrating the predictions of all hypotheses might not – for reasons of computational efficiency – be implemented in either minds or brains in a direct, straightforward fashion. One possibility is that human learners could approximate the computations of Bayesian inference through the judicious application of a MIN rule-based heuristic and a flexible similarity-based heuristic (which is sensitive to the number and distribution of examples in the way that Bayesian inference requires). Because they would be specialized for different regimes of concept learning, such rule-based and similarity-based heuristics might be implemented in different brain modules (as some preliminary evidence suggests (Smith et al., 1998)). Also, as we saw in the number game, a simple rule found to be consistent with a small number of examples takes precedence over similarity-based generalization – both in the Bayesian analysis and in people's behavior. Hence, if rule-based and similarity-based generalization are the responsibilities of separate brain modules, we might expect to see the rule module exerting an inhibitory influence over the similarity module, in order to implement this computational preference.

Obviously, I am only speculating about the possibilities of neural architecture here. The key point is that the Bayesian framework, while it is an account at the level of competence (Chomsky, 1986) or computational theory (Marr, 1980), does give some insight into questions of how the ability to learn concepts from examples might be implemented in the mind/brain: why concept learning processes might need to be modular (to achieve computational tractability); why there might be two modules, one best described as rule-based and the other best described as similarity-based (to capture the two extremes of graded and all-or-none generalization behavior); how each individual module should respond to a set of examples (weighting candidate rules or features of similarity according to the size principle); and how the modules should interact to accomplish a unified computational goal (with the similarity module dominating initially and the rule module coming to dominate as more examples are observed, at a rate that depends on the density of hypothesis overlap.)

### 6.2.5 The interaction of prior knowledge and observed examples in concept learning

The importance of prior knowledge and intuitive domain theories in guiding the acquisition of natural concepts has been a central theme in the cognitive literature for the last 15 years (Murphy & Medin, 1985; Carey, 1985). However, the formal models of concept learning currently popular among mathematical psychologists – exemplar models and connectionist models – are for the most part unable to incorporate structured domain knowledge in a natural way and focus instead on the statistical information provided by the observed examples. In particular, formal models of concept learning are applied almost exclusively to tasks that use artifical stimuli and that provide both positive and negative examples provided to the learner, two factors which significantly reduce the demand for strong *a priori* constraints on generalization. Partially as a result of the divide between formal modelers and other psychologists studying concepts, much effort has been spent in debating whether prior knowledge or observed examples provide the more important force behind the acquisition of con-

cepts (Jones & Smith, 1993 ). This version of the "nature versus nurture" debate threatens to miss what makes human concept learning so remarkable: the *interaction* of these two forces, which accomplishes what neither one could on its own.

The essence of the Bayesian framework of this thesis lies in capturing this interaction. The mathematics here could not be simpler. The posterior probability of a hypothesis is proportional to the product of its prior probability and its likelihood given the examples:

$$p(h|X) \propto p(X|h)p(h).$$

The likelihood captures the force of the data, giving preference to hypotheses under which the examples are more likely to be observed. The prior captures our beliefs about the relative naturalness of different candidate extensions for a concept – independent of the particular examples observed. Combining these terms via a product is signficant when we are dealing with probabilities. It means that (1) Bayes thinks these two sources of knowledge should be combined conjunctively, as in an AND operation, and (2) Bayes treats them as if they are independent variables. At an abstract level, both of these features seem like they have to be correct. Whether a candidate extension for a concept is plausible a priori is independent of its statistical support in the observed examples. And for a hypothesis to be accepted, it must be conceptually natural and be supported by the examples observed. Whether this structure describes what actually goes on inside people's heads is an empirical matter, which the three case studies were meant to address.

In all of the case studies, we saw that neither prior knowledge nor the observed data could alone claim any responsibility for how people learned and generalized from only one or a few examples. The Bayesian models we developed were able to incorporate the relevant domain knowledge that participants in these studies might reasonably be assumed to have, in the form of a hypothesis space of candidate extensions and a prior distribution over that space, and to show exactly how that knowledge served to guide generalization from the examples provided. In the healthy levels domain of Chapter 3, the prior knowledge that the substance in question was an environmental pollutant

(*e.g.* lead), as opposed to a naturally produced bodily chemical (*e.g.* cholesterol or insulin), had a dramatic on how people generalize from just one example, which the Bayesian framework was able to capture by restricting the hypothesis space to include only intervals of healthy levels with their minimum at 0. In the word learning domain of Chapter 4, the hypothesis space was given a nested tree structure inspired by the taxonomic bias for learning nouns (Markman, 1989). The model's predictions were also made significantly more accurate when the prior probability included a bias for mapping words onto basic-level classes (Markman, 1989; Mervis & Crisafi, 1982). In the number game of Chapter 5, the hypothesis space incorporated multiple kinds of prior knowledge, including both formal knowledge about mathematically priveleged classes (*powers of two*, *even numbers*) as well as less formal knowledge about numerical magnitude.

In none of these cases was prior knowledge about the possible extensions of concepts alone sufficient to explain the course of concept learning. The examples exerted great statistical power through the size principle embodied in the likelihood term:

$$p(X|h) = \left[\frac{1}{\text{size}(h)}\right]^n$$

if $X \in h$, and 0 otherwise. In each case study, the size principle determined how far people generalized as a function of the number of examples observed. Increasing numbers of examples led to increasingly conservative bounds of generalization, always converging to the smallest hypothesis consistent with the observed examples. The size principle also determined in what ways people generalized, causing the gradients of generalization in the continuous feature space of the healthy levels task to shrink along dimensions with tighter clustering of examples, or, in the number game, causing a specific mathematical rule, *e.g. powers of two*, to take precendence over hypotheses based on numerical magnitude, *e.g. numbers less than 80*".

Most intriguingly, both the structure of the hypothesis space and the size principle were essential in characterizing the difference between domains that give rise to similarity-based versus rule-based generalization as the default mode, and for un-

derstanding domains like the number game in which both modes coexist. Whether people's hypothesis space has a sparse or densely overlapping structure is determined by their prior beliefs about the possible concepts in a particular domain, but the reason why this structural difference leads to a difference in generalization behavior after just a few examples lies in the statistical force of the examples, as expressed through the size principle.

In sum, I hope that by providing a framework for understanding the interaction of domain-specific prior knowledge and general-purpose statistical principles, the Bayesian approach will help to bring formal models of human concept learning back from their long-standing focus on artificial, knowledge-poor tasks and into closer touch with the more natural, knowledge-intensive learning settings that pose the real challenges for cognitive science.

## 6.3 Other directions

A number of extensions, challenges, and broader applications of the Bayesian framework are beyond the scope of this thesis to cover in detail, but I will try to sketch some of the important points of ongoing and future work in this section.

### 6.3.1 The complexities of learning in the real world

Real-world learning situations present a number of complexities which I have not addressed here, but which the Bayesian framework can be extended to handle. Some of these issues include noisy example values, noisy example labels, disjunctive concepts (with two or more possible extensions) and weak prior knowledge. The Bayesian framework deals with these added complexities by supplementing the generative model with extra hidden variables in addition to the hypothesis space of candidate extensions for the concept. These hidden variables can denote, for instance, which examples are labeled correctly, which examples are drawn from the same (of two or more) extensions, or which features from a large basis set are relevant for defining the concept. The learner can assign probabilities to these hidden variables just as

he does to the candidate extensions, using a prior probability combined with a size-based likelihood. The results are automatic procedures for assigning error margins to example values, rejecting potential outliers, controlling the complexity of the inferred extensions, and selecting the relevant features. Appendix C explains the ideas behind these more subtle Bayesian inferences in more detail and illustrates the generalization gradients they give rise to, using the rectangular regions hypothesis space from Chapter 3 as a base.

### 6.3.2 Challenges for Bayesian inference

An important theme of this thesis has been to draw connections between certain qualitative properties of the Bayesian formalism and important qualitative aspects of human generalization behavior. However, there are other aspects of Bayesian inference which we did not dwell upon and which do not seem to map so well onto human behavior. One major point of difference is that the behavior of most Bayesian models, including all models in this thesis, is independent of the order in which examples are encountered, while it seems unlikely that this will be true in general for human learners (*e.g.* Elio & Anderson, 1984; Goldman, 1986). However, issues like this one are not refutations of the Bayesian framework, but challenges to it. Bayesian inference is only insensitive to the order of experience under the assumptions that each example is an independent sample from the concept, and that the relevant sampling probabilities are not changing over time. Neither of these assumptions is strictly true in many situations, but they make the theory much simpler to work with. In future work, it will be important to develop Bayesian models which do not make these independence and stationarity assumptions and to see how well they describe the effects of example order on human generalization behavior.

Another important challenge for Bayesian models is the problem of learning when the true concept is not in the learner's hypothesis space. All guarantees of optimality or rationality are off in these circumstances, which may arise frequently for real-world learners. In future work, I would like to understand how Bayesian models of concept learning generalize in these situations, in particular, relative to human learners. Peo-

ple have a remarkable ability to enlarge their hypothesis space in apt ways when the observed data are not consistent with any hypothesis that they are currently entertaining – "necessity is the mother of invention" – and trying to understand these "invention"-like processes within a Bayesian framework, perhaps using the techniques discussed in the previous subsection and in Appendix C, is another goal for future work.

### 6.3.3 Implications for machine learning

All along, I have taken the position that the capacities of human concept learning far outstrip those of any artificial system, and thus that we can only make progress on understanding human learning by developing more sophisticated computational models than are currently available in the machine learning literature. Having made at least some progress on understanding how people can generalize from just a few positive examples, it now makes sense to turn back to machine learning and explore the implications of this work for building better artificial concept learners. This is mostly a topic for future research, but we can identify a few important principles based on the results of this thesis.

First, when designing statistical algorithms for concept learning, it is crucial to have the right probabilistic model for how the examples are generated. Many machine concept learning systems assume what I called "weak sampling" in Chapter 2, where the examples are sampled independently from the concept to be learned. This is an appropriate assumption for many discrimination learning tasks, *e.g.* when the computer has to learn to classify manufactured parts into one of two mutually exclusive classes. But it is not appropriate for learning concepts from positive examples that a human user provides, and it does not lead to the size principle that was responsible for rapid convergence to the true concept. The size principle requires something like the "strong sampling" model, in which the (positive) examples are assumed to be sampled from the concept to be learned. An immediate goal for future work is to build a computer system that learns concepts from positive examples provided by a human user (and perhaps also from negative examples provided as feedback on the

system's mistaken generalizations), and to explore the impact of using an appropriate generative model (strong sampling, as opposed to weak sampling) on how natural and easy to use human users find the system.

Second, in designing machine concept learning systems a great deal of attention should be paid to the structure of the hypothesis space. In this thesis, we saw how certain structural features of the hypothesis space determined important features of learning behavior, *e.g.* how the density of hypothesis overlap determined how fast human learners (and Bayesian models) converged on the true concept. If very rapid learning is goal for a machine learner, then it is essential to have a hypothesis space with sparse overlap structure. Engineering hypothesis spaces with theoretical desirable properties thus turns out to be one of the most important problems in designing an artificial concept learning system.

The final message of this work for machine learning is that we shouldn't be so afraid of jumping to conclusions. As we saw in Chapters 4 and 5, human concept learners frequently jump to conclusions with no logically sound basis after very few examples. The Strong Bayes concept learning framework also shows this kind of behavior, in contrast to traditional machine learning theories (Valiant, 1984; Haussler, 1988; Kearns & Vazirani, 1994; Vapnik, 1995) which aim for asymptotic performance guarantees at the cost of learning curves that are, by human standards, hopelessly slow. As I wrote in Chapter 3, if we want our learning algorithms to be able to learn concepts from just a few examples, the way that people do, then we have to be willing to accept that sometimes they will leap to incorrect generalizations, just as people do. Bayes takes the bad aspects of human learning along with the good; conventional theories of machine learning take neither.

### 6.3.4 Where do the priors come from?

The Bayesian framework of this thesis gains much of its explanatory power from assumptions about the learner's hypothesis space and prior probability distribution over that space. This naturally raises the question of where these priors come from, both for the learner who must use them to generalize and for the experimenter who

must use them to model the learner's generalization behavior. I am not ashamed to admit that I have no simple answer to this question. In fact, I think that this is a strength of the theory, not a flaw. People bring many different kinds of knowledge to bear on concept learning tasks, which may vary dramatically from domain to domain. In order to describe human learning behavior in a wide range of situations, a theory must be able to accomodate many different kinds of prior knowledge. Explaining the origin of this prior knowledge is an important part of explaining the origin of our concepts, but it is in an important sense not the proper subject for a theory of concept learning. A theory of concept learning, such as the Bayesian framework of this thesis, is properly about the *mapping* from prior knowledge to generalization behavior. This does not mean that the Bayesian framework is of no help in understanding the nature of people's prior knowledge. Quite the contrary, one way to view the Bayesian framework is as a tool for studying prior knowledge, in particular, for making "forward" predictions about the effects of certain structural features of prior knowledge (*e.g.* density of hypothesis overlap) on behavior, as well as for making "backwards" inferences about the structure of people's prior knowledge from the kinds of generalization behavior they produce.

Even though there is no simple answer to the question of where the concept learner's prior knowledge comes from, two possible sources seem like good candidates: unsupervised learning and domain theories. The hypothesis spaces used in this thesis either consisted of low-dimensional feature spaces (Chapter 3), clusters of objects (Chapter 4), or a mixture of the two (Chapter 5). These are exactly the sort of structures that many unsupervised learning algorithms are designed to discover in raw data (Duda & Hart, 1973; Shepard & Arabie, 1979; Tenenbaum, 1995, 1998). The idea of using unsupervised learning to provide the raw material for supervised learning is a standard one in the machine learning tradition (Bishop, 1995). In fact, the clustering algorithms I used in Chapters 4 and 5 to construct the hypothesis spaces are essentially unsupervised learning algorithms, which operated on subjects' similarity data as their measurements of objects. Exactly the same algorithms could be used on raw images or other perceptual signals, given an appropriate similarity

measure (which again will depend on the domain and the learner's goals).

Domain theories are another important source of *a priori* constraint on concept learning. In some cases, a domain theory will generate the hypothesis space directly, by suggesting particular subsets of objects as important candidates for the extensions of new concepts. In other cases, a domain theory affects the hypothesis space indirectly, *e.g.* by dictating a particular kind of structure which can then be detected by an unsupervised learning algorithm. The full hypothesis space would then arise from the interaction of the domain theory with the unsupervised learning algorithm. This was the case in the word learning studies of Chapter 4, where the intuitive notion of a tree-like taxonomy of object kinds, combined with the taxonomic bias of word learners, suggested a hierarchical clustering algorithm for generating the learner's hypothesis space.

The Bayesian framework for concept learning also gives us a set of criteria for evaluating new proposals for unsupervised learning algorithms or domain theories, in terms of how well they mesh with the framework's needs and constraints, and in terms of what kind of learning behavior they lead to under Bayesian inference. For example, the fact that only sparse hypothesis structures lead to very rapid concept learning in the Bayesian framework may imply that, in a given domain where people typically learn all-or-none concepts from just a few examples, we should focus our attention on those domain theories or clustering algorithms which lead to sparsely structured hypothesis spaces for learning.

### 6.3.5 What makes good examples of a concept?

Phenomena of "typicality", "representativeness" or "exemplar goodness" are some of the most robust in the cognitive literature, but also some of the most poorly defined (Kahneman & Tversky, 1972; Rosch & Mervis, 1975; Barsalou, 1985; Gigerenzer, 1996). The framework of this thesis suggests a precise definition of one sense of representativeness, in the context of Bayesian concept learning. That is: a good example of a concept is one which leads a Bayesian learner to generalize to all (or most) of the other entities in that concept, and only those entities. This is closely related to

Feldman's (1997) idea of a logically generic example, although more general because of its probabilistic formulation. Most importantly, the probabilistic formulation allows us to define representativeness for a set of examples – what makes $x_1, \ldots, x_n$ a good sample of a concept – which is important if we want to understand how multiple examples interact to aid concept learning. This Bayesian definition of representativeness can be used to resolve a number of puzzles about typicality. This will be the subject of a forthcoming paper, but the basic idea is summed up in the following sketch (adapted from Tenenbaum, 1997a).

**Genericity: a Bayesian definition of representativeness.** Let $\mathcal{H} = \{H_1, \ldots, H_n\}$ denote a set of mutually exclusive hypotheses that might account for an observation $D$. Then Bayes' rule asserts that $D$ supports $H_i$ (maximizes $p(H_i|D)/p(H_i)$) to the extent that the occurence of $D$ is better explained (more probable) under $H_i$ then under any of the other alternative hypotheses $H_{j \neq i}$, weighted by their priors. An observation $D$ that provides strong support for $H_i$ under this measure will be called a *generic* example of $H_i$. Of the many ways in which an outcome or object may be typical of a process or category, being a good example for the purposes of inductive generalization is surely one of the most natural. In contrast to many previous proposals, this sense of typicality as genericity is quite precise and follows from normative principles of inductive inference. It also clarifies a number of otherwise puzzling phenomena:

**1. Typicality may not imply typical features.** Under the standard view, people judge a robin to be a typical bird because it shares many salient features with the bird prototype, or with other birds. But people judge $S_1 =$ `HHTHTTTH` to be a typical sequence of fair coin flips not because it has certain salient features, but because it does *not* have the features (patterns or biased tendencies) that less typical sequences like $S_2 =$ `HTHTHTHT` or $S_3 =$ `HHTHTHHH` do. On a Bayesian analysis, those salient features make $S_2$ and $S_3$ less generic, and thus the "featureless" $S_1$ is correctly identified as the most typical.

**2. Typical individuals may not belong to typical subclasses.** Robbie the robin would be judged to be a typical or representative bird, and robins are

also considered to be a typical *kind* of bird. But consider the two quadrilaterals in Fig. 1A and Fig. 1B. Most people consider 1A to be more typical or representative of quadrilaterals. Yet 1B, as a rectangle, belongs to a typical kind of quadrilateral, while 1A does not. In fact, stimuli belonging to salient subclasses of a category are less generic, and thus less likely to be seen as representative members of the category.

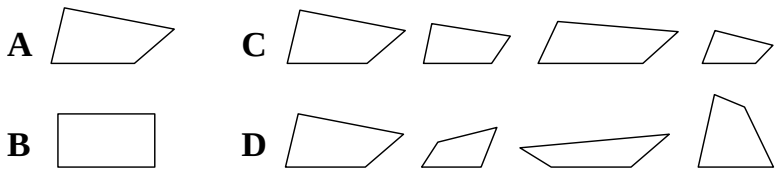



**Figure 1**

**3. Typicality may be compatible with definitional categories.** It is well known that even apparently definitional categories such as "odd number" exhibit reliable typicality gradients (Armstrong et al., 1983). These results suggest that there is no simple relation between typicality effects and category structure. But definitional categories and typicality phenomena may coexist peacefully: Fig. 1B clearly satisfies the definition of a quadrilateral, but as a rectangle, it is not a generic quadrilateral like Fig. 1A. Typicality gradients determined by genericity may tell us nothing about the structure of any one category on its own, but a great deal about the structure of the set of categories that comprise the hypothesis space for inductive inference.

**4. "A set of typical X's" may not equal "a typical set of X's".** In general, categories are best learned from experience with typical members (Mervis & Rosch, 1981). Since category learning usually involves experience with more than one example, the concept of "a typical X" should be naturally extendable to "a typical sample of X's". But this is not trivial: Fig. 1A is a representative quadrilateral, while a sequence of similar shapes (Fig. 1C) is not a representative sample of quadrilaterals. However, under the assumption of independent sampling, the same notion of genericity that picks out typical quadrilaterals (Fig. 1A) and typical sequences ($S_1$) also

distinguishes typical sequences of quadrilaterals (Fig. 1D).

### 6.3.6 Reasoning with categories

Some of the most famous fallacies of "irrational" reasoning depend on categorization and concept learning. These include the conjunction fallacy, or "Linda effect" (Tversky & Kahneman, 1983), and Wason's "2-4-6" task (Wason, 1960). However, while the typical person's behavior in these tasks is unquestionably far from the experimenter's definition of the rational or normative response, it is not without its rational behavior. In fact, if these tasks are viewed as tasks of learning concepts from limited positive evidence, then people's behavior makes a great deal of sense in light of the Bayesian theory of this thesis. This will be the subject of a future paper, but the basic idea can be summed up in a sentence by reference to the size principle. In both the "Linda" and "2-4-6" tasks, subjects choose a category that "too small" (according to the experimenter's defintion) – "feminist bankteller" as opposed to "bankteller"; "three numbers increasing in steps of two" as opposed to "any three increasing numbers" – but this is exactly what we should expect if people are reasoning in accordance with the size principle.

Not only intuitive reasoning, but scientific reasoning as well, shows the stamp of Bayesian concept learning. Popper's (1959) argument that confirmation in scientific inference does not follow the laws of probability is directly analogous to Tversky and Kahneman's (1983) argument that people are not Bayesian in the Linda task. Just as subjects' performance in the Linda task may be described as Bayesian under the appropriate model of concept learning, so can scientific inference can be described in Bayesian terms under the appropriate formulation of the inference problem. The idea that scientists choose the "most powerful" or "strongest" theory (what Popper (1959) calls the theory with highest "corroboration") is simply the same as saying that they choose the most probable theory under the size principle. Also, the Bayesian theory's account of "what makes a good sample of a concept?" (discussed above) is directly related to Bayesian justifications of the preference for diverse evidence in scientific inference (Horwich, 1982). In future work, I hope to further explore the parallels

between mechanisms of generalization in scientific reasoning versus those underlying everyday cognition.

## 6.4 Are people "really" Bayesian?

One of the most vigorous debates in cognitive psychology centers on the question of whether or not it is generally true that people reason in accordance with the normative prescriptions of Bayesian inference. The classic work of Tversky and Kahneman is well known for its findings of broad and blatant violations of probability theory in human reasoning (Tversky & Kahneman, 1983; Kahneman & Tversky, 1972). More recently, psychologists inspired by evolutionary and computational considerations have argued that human reasoning does in fact embody Bayes' theorem at its core; however, we may only be able to observe this experimentally when subjects are placed in evolutionarily realistic scenarios with stimuli represented in computationally felicitous ways (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1996; Brase, Cosmides & Tooby, 1998).

Because I have suggested that concept learning, and perhaps other phenomena of human reasoning as well, may be understood in terms of Bayesian inference, I have some obligation to comment on how my work fits into this debate. I have two comments to make. First, the classic debate about whether or not "people are Bayesian" focuses primarily on whether or not Bayes provides a *quantitatively* accurate model of human judgment. Specifically, when people give judgments of probability, do they obey the normative prescriptions of probability theory or not? In contrast, my interest in Bayesian models is primarily for their *qualitative* properties. To quote Glenn Shafer (via Judea Pearl), "Probability is not really about numbers; it is about the structure of reasoning." The Bayesian framework of this thesis is a model of what the core computational structure of human concept learning must be like: what the several essential components are, how they interact, and what are the consequences of ignoring one of them. Of course, I would like to make quantitative predictions of human behavior as much as the next psychologist. But that is a

relatively minor goal compared to answering questions like the following. How can concepts be learned from only one or a few positive examples? What kind of prior knowledge do we bring to bear on a concept learning task? How does that knowledge interact with the examples we observe to guide our generalizations? When and why is generalization based on rules versus similarity, or some combination of both?

My second comment is this. The heuristics and biases debate is primarily about whether or not "people are Bayesian" in some general-purpose sense. In contrast, I have argued for a Bayesian theory to explain a particular human ability, the ability to learn and generalize object concepts from just a few positive examples. I have in this final chapter suggested that other phenomena of reasoning may be understood in this framework, but only to the extent that they draw on our natural and distinct capacity for concept learning. I have specifically *not* argued that Bayesian inference provides the best way to describe human thinking or reasoning in general. I would not argue against the general-purpose claim either; rather, I think we must for the moment remain agnostic, and for good evolutionary reasons.

The brain has evolved mechanisms for solving particular computational problems, and the problem of learning and generalizing concepts from very limited evidence is almost certainly one of those. Other such problems occur in visual perception, motor control, sentence processing, and language acquisition, to name only a few. I do not mean to suggest that concept learning is as modular as some of those other abilities appear to be, only that these are all more or less well-specified computational problems for which we can begin to formulate and test the necessary ingredients of a Bayesian theory: the hypothesis space, the data, the prior and the likelihood. Indeed, Bayesian models are now the state-of-the-art for many of these problems, in both the psychological and computational literature (Weiss, 1998; Todorov, 1998; Narayanan & Jurafsky, 1998; Brent & Cartwright, 1996). But none of these theories about vision or motor control or sentence processing makes any claims about whether or not "people are Bayesian" in general! None of them *could* make such claims; they are not theories about people in general, but about specific computational problems that people solve. My Bayesian framework for concept learning falls into the same class. While clearly

closer to central, domain-general cognitive processes than is vision or motor control, learning concepts from examples is nonethless a distinct ability focused on solving a particular set of computational problems. A theory of concept learning, to the extent that it addresses those particular problems, says nothing about the question of whether or not "people are Bayesian" in some general-purpose sense. There, we must remain agnostic until we have a better understanding of the computational problems involved in common-sense reasoning under uncertainty, until we have some idea of what a serious general-purpose theory of cognition – Bayesian or otherwise – would even look like.

Of course, for those who doubt that there is such a thing as a "general-purpose theory of cognition" in any form (Fodor, 1983), the question of whether or not "people are Bayesians" doesn't even make sense. But to those skeptics, I say: wait and see. Fodor (1983) arrived at his skepticism of general-purpose cognitive theories by analogy with the impossibility of a general-purpose theory of confirmation in philosophy of science. However, increasing numbers of philosophers are now coming to believe that the Bayesian view offers deep general insights into the nature of inductive confirmation in science (Earman, 1992; Howson & Urbach, 1989; Forster, in press), and truly clarifies many long-standing puzzles (Watanabe, 1960; Horwich, 1982, 1993). Bayes is not (yet) in any sense a "general-purpose theory of confirmation", but how far it will take philosophy of science is still an open question that many approach with guarded optimism (Earman, 1992).

To update Fodor's analogy, I think we can say much the same for the role of Bayesian theories in cognitive science. I have tried to show here how the framework of Bayesian inference illuminates the core computational structure of one particular human cognitive ability, that of learning concepts from examples. Many heated contemporary debates about the nature of concepts – is conceptual knowledge primarily based on rules or similarity? is generalization driven primarily by prior knowledge or observed data? – either dissolve or yield real ground on a Bayesian analysis. Other phenomena of reasoning which are not usually thought of in terms of concept learning also seem to make more sense in light of the Bayesian theory. Whether these results

will ultimately lead to progress on a "general-purpose theory of cognition" – whether that even makes sense as a goal – is yet to be determined. Regardless, I think we are far from seeing the full impact of Bayes on the study of the mind.

# Appendix A

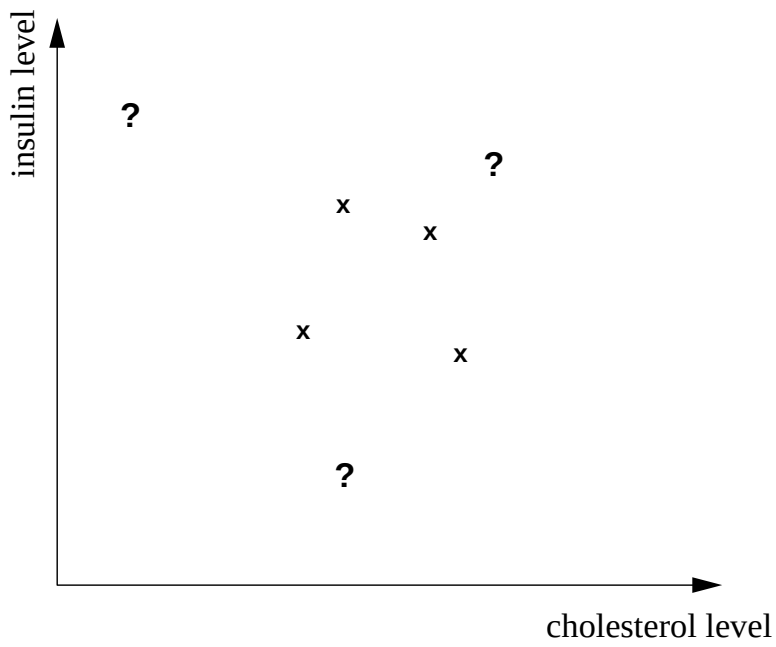
# Why standard models of discrimination learning are not appropriate as general models of human concept learning

The theoretical study of concept learning has been at the center of both cognitive psychology and machine learning since these fields' inceptions (Bruner, Goodnow & Austin, 1956; Shepard, Hovland & Jenkins, 1961; Hunt, 1962; Mitchell, 1979). Yet despite the long tradition of formal learning models in both fields, previous work has not, for the most part, addressed the questions I consider in this thesis of how far and in what ways to generalize a concept from only a few positive examples. Mathematical modeling in both cognitive psychology and machine learning has focused extensively on *discrimination learning* tasks, which look similar to the concept learning tasks considered in this thesis but differ from them crucially in giving positive and negative examples equal and essential roles in guiding generalization. As a consequence, models developed to account for discrimination learning *require* negative examples of a concept in order to generalize in any meaningful way. This is quite unlike people (or the Bayesian models developed in this thesis) who, in all of our experiments and

in their daily lives, are capable of generalizing meaningfully – if not always accurately – from strictly positive evidence. This appendix reviews in some detail the armory of models developed for discrimination learning and illustrates why they cannot account for how people learn concepts from just a few examples.

In discrimination learning tasks, the learner is given examples of both positive and negative instances of a concept and is required to learn some procedure for discriminating the positives from the negatives. Here's an example in the number concept domain: four numbers the program accepts are 16, 8, 2, and 64; four numbers the program does *not* accept are 41, 10, 13 and 98. Now which other numbers do you think the program will accept? Some discrimination learning tasks are posed as *classification* tasks; instead of seeing positive and negative examples of a single concept, the learner receives positive examples of two or more classes *that are assumed to be mutually exclusive*. For instance, I could describe the physical traits of a set of martian animals and teach you to classify these animals into one of several different species by giving you several examples of each species. Because the species are assumed to be mutually exclusive, the positive examples of one class are implicit negative examples for all the other classes. Thus classification tasks like this one are also instances of discrimination learning; the learner's job is to figure out how to discriminate the positive instances of each class from all of the other objects.

There have been many empirical studies of human discrimination and classification learning (Bower & Trabasso, 1964; Fried & Holyoak, 1984; Gluck & Bower, 1988), and the pattern recognition and machine learning literatures are full of well-known techniques for solving these learning problems (Duda & Hart, 1973; Mitchell, 1997; see also Nosofsky (1992), Ashby (1992), and Ashby & Leola-Reese (1995) for versions of these models in the psychology literature). All of these tecniques require some *input representation*, some way of representing the objects to be classified in terms of the properties that will be relevant for the classification procedure. The right choice of properties to represent will obviously vary from domain to domain and task to task. In the number concept game, we might want to represent both the mathematical features of numbers – *even*, *prime*, or *power of 3* – as well as their magnitude features




"healthy levels"

**Figure 1**

– *between 30 and 45*, *less than 10*, etc. In many standard discrimination settings, objects are represented as points in a continuous multidimensional space, where the axes of the space correspond to the relevant continuous-valued features. For example, if we wanted to learn to tell the difference between healthy people and unhealthy people based on their blood samples, we might measure the levels of cholesterol, insulin, and other substances in their blood, and then represent each person as a point in cholesterol-insulin space (Figure 1). We'll stick with this simple two-dimensional feature space example for this section, because it is the most common setting in which discrimination and classification tasks have been studied.

There are two main paradigms for modeling discrimination learning. *Direct* approaches learn a direct mapping $f : \mathbf{v} \rightarrow l$, from input features $\mathbf{v}$ (*i.e.* $v_1$ = cholesterol level, $v_2$ = insulin level) to output labels $l$ (*i.e.* l = { *healthy* / *unhealthy* }). The discriminant function $f$ is optimized over the training set of positive and negative examples to map, as closely as possible, each example's input features to the ap-

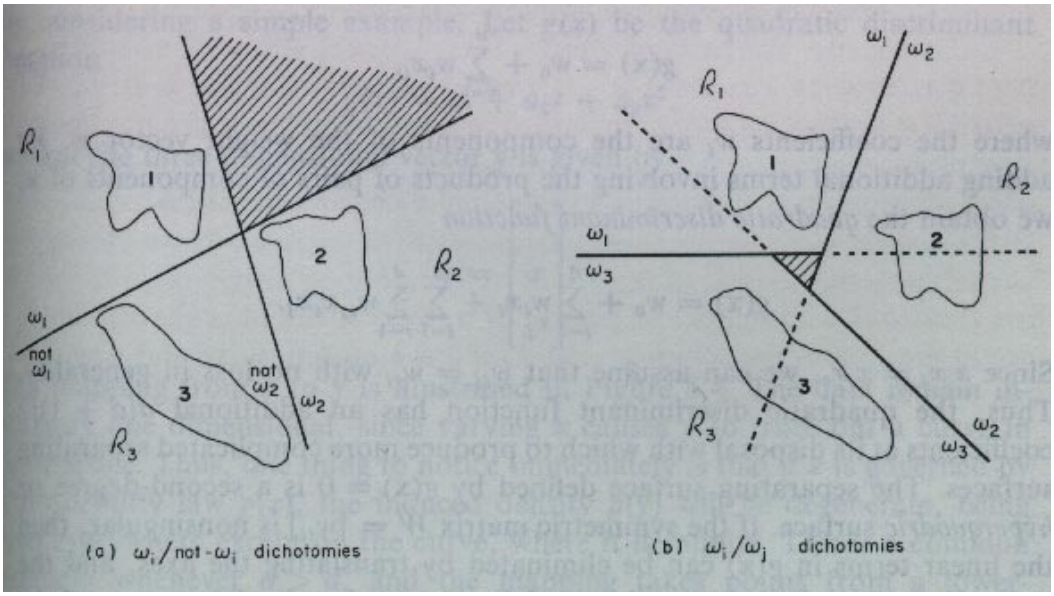
propriate output label. Standard direct approaches include many neural network architectures (single- and multi-layer perceptrons (Gluck & Bower, 1988; Shanks, 1988; Estes, 1994); radial basis function (RBF) networks (Sung & Poggio, 1994; Poggio & Girosi, 1990); and hybrid exemplar-based networks Kruschke, 1992; Aha & Goldstone, 1992) and many statistical pattern recognition algorithms (linear discriminant analysis; general recognition theory (Ashby, 1992); support vector machines (Vapnik, 1995)).[1] Figure 2a (from Duda & Hart, 1973) illustrates a direct model that uses linear discriminant functions to solve a two-class classification problem in a two-dimensional feature space like the healthy levelsspace.

Direct approaches learn *no* model of a concept, but only a way to discriminate positive from negative instances. *Indirect* or *model-based* approaches learn a model of the positive examples and another model of the negative examples; they subsequently classify instances as positive or negative according to which model fits better.

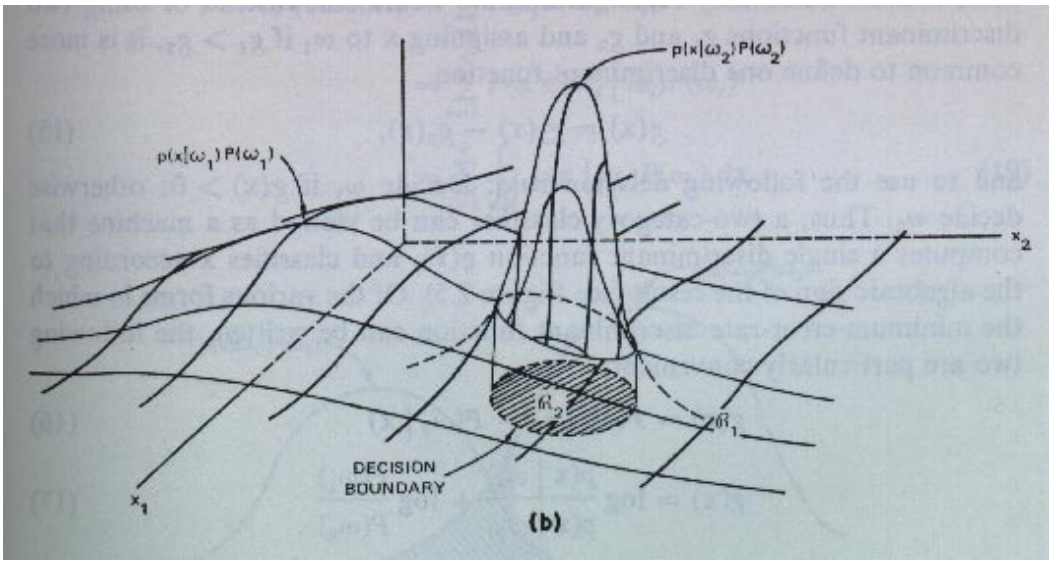
The most common (and most principled) indirect methods are *distributional* approaches, which model a concept as a probability distribution over some space of features $\mathbf{v}$ (or probability density, in a continuous feature space) and classify new instances $x$ as members of $C$ according to their estimated probability $p(\mathbf{v}(x)|C)$. The simplest form for the density $p(\mathbf{v}|C)$ might be a gaussian distribution (Fried & Holyoak, 1984; Duda & Hart, 1973), but in general $p(\mathbf{v}|C)$ may be represented in many different forms: mixtures of gaussians (Bishop, 1995); nonparametric density models, both kernel-based (Ashby & Leola-Reese, 1995) and network-based (Hinton et al., 1994; Jaakola et al., 1997); and loosely speaking, "autoencoder" neural networks (Gluck & Myers, 1993; Japkowicz, Myers & Gluck, 1995; Petsche et al., 1997). In order to classify new objects as positive or negative instances of $C$, we also need to learn a distributional model of the negative examples, $p(\mathbf{v}| \sim C)$, and the base rate of positive examples, $p(C)$.

The function used to classify new objects is the class posterior probability $p(C|\mathbf{v})$,

[1]The "nearest neighbor" family of classification techniques (Cover & Hart, 1967) may also be placed in this category, although they implicitly embody a density model (with asymptotic error rate no worse than twice the optimal density model) that places them close to nonparametric distributional approaches (see below).




**A *discriminative* approach to classification**




**A *distributional* approach to classification**

(from Duda & Hart, 1973, pp. 19, 133)

**Figure 2**

the probability of belonging to $C$ given the observed feature values $\mathbf{v}$, which we compute from $p(\mathbf{v}|C)$, $p(\mathbf{v}| \sim C)$, and $p(C)$ via Bayes' theorem: [2]

$$p(C|\mathbf{v}) = \frac{p(\mathbf{v}|C)p(C)}{p(\mathbf{v}|C)p(C) + p(\mathbf{v}| \sim C)(1 - p(C))}. \tag{A.1}$$

Figure 2b (also from Duda & Hart, 1973) depicts a Bayes classifier over a two-dimensional feature space.

Direct and distributional approaches to discrimination share a common handicap, which rules them out for the purposes of this thesis: they cannot generalize in a principled way from only positive examples. For direct approaches, this point is obvious. We cannot directly optimize a function that discriminates positive and negative instances if we have seen only examples of the positives! For distributional approaches the point is more subtle. Couldn't we just assume some reasonable default density for the negative examples, and some reasonable default base rate $p(C)$, and plug this into the Bayes' classifier to obtain an estimate of the posterior $p(C|\mathbf{v})$?

Figure 3 shows the dangers behind this assumption, for the simple case when the examples of a category are distributed with a Gaussian density. Each row in Figure 3 depicts the same three positive examples, along with a different set of negative examples. The left column shows the Gaussian densities $p(\mathbf{v}|C)$ and $p(\mathbf{v}| \sim C)$ for the positive and negative examples respectively. The right column shows the resulting posteriors $p(C|\mathbf{v})$ computed from Equation A.1, which determine how the algorithm generalizes the concept. Observe that a single set of positive examples can give rise to very different looking patterns of generalization depending on the negative examples observed. In some cases, the posterior $p(C|\mathbf{v})$ has the same qualitative form as the distribution of positive examples $p(\mathbf{v}|C)$ (Figure 3, last two rows), but in other cases it looks quite different. Clearly, there is no single default assumption for $p(\mathbf{v}| \sim C)$ that will serve even moderately well for all possible sets of negative examples that could be observed.

[2] Don't confuse this with Bayesian concept learning. Both use Bayes' theorem, but that's where the similarity ends. The real difference is in how they represent a concept....

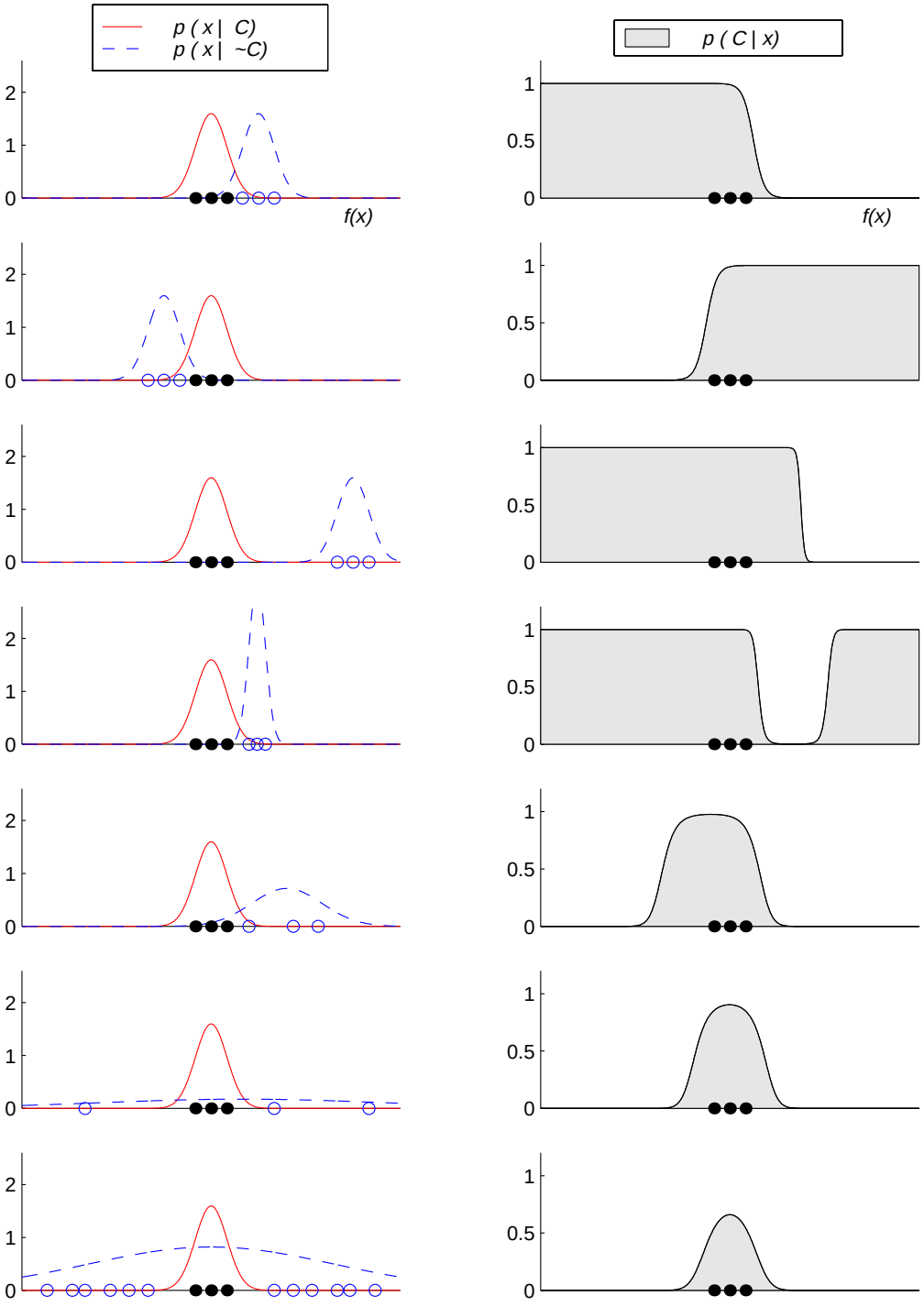



**Figure 3**

At this point, we might wonder whether we really need negative examples at all in a distributional approach. Why don't we just use $p(\mathbf{v}|C)$ directly to classify new objects? This is the idea behind an alternative distributional technique known as "novelty detection". We choose some threshold $\theta$ on $p(\mathbf{v}|C)$, and then classify only those observations $\mathbf{v}$ with $p(\mathbf{v}|C) > \theta$ as positive instances of $C$; all others are rejected as "novel" (*i.e.* negative instances). The problem here is how to choose $\theta$. As Figure 4 illustrates, any particular value of $\theta$ (equal to 0.1 in this figure) will lead to reasonably intuitive generalization behavior for some distributions (column two), but not for distributions with much larger or smaller ranges of variation (columns one, three, or four).

We could try to solve this problem by setting $\theta$ to classify the top 95% (or some other reasonable percentage) of the probability mass in $p(\mathbf{v}|C)$ as positive. For a gaussian distribution in one dimension, with variance $\sigma^2$, this criterion includes all values within $1.96\sigma$ of the mean. Such a criterion is a standard choice for setting confidence intervals on estimates of population parameters, such as the mean of a distribution. However, our primary interest here is not to estimate the mean class member, but rather to identify positive instances of the concept! As a concept learning algorithm, thresholding on the top 95% of the density $p(\mathbf{v}|C)$ fails to meet a basic desideratum of asymptotic consistency. That is, we would hope that as the number of examples observed approaches infinity, our concept should approach the true concept and the number of classification mistakes should approach zero. But if we set $\theta$ to classify exactly 95% of the positive instances as positive, we will *always* mistakenly reject 5% of the true instances of this concept, including 5% of the labeled examples *known to be positive!* At the same time that we reject explicitly labeled positive examples, we will also accept as positives some novel objects which we have never seen labeled as such. This sort of behavior might be appropriate if we expect that our observations are noisy (either in their observed feature values or in their labels), but it is clearly unreasonable when observation noise is negligible or nonexistent. When we are 100% confident that the positive examples provided really do belong to the concept, as in the number concepttask, stimuli identical to those examples

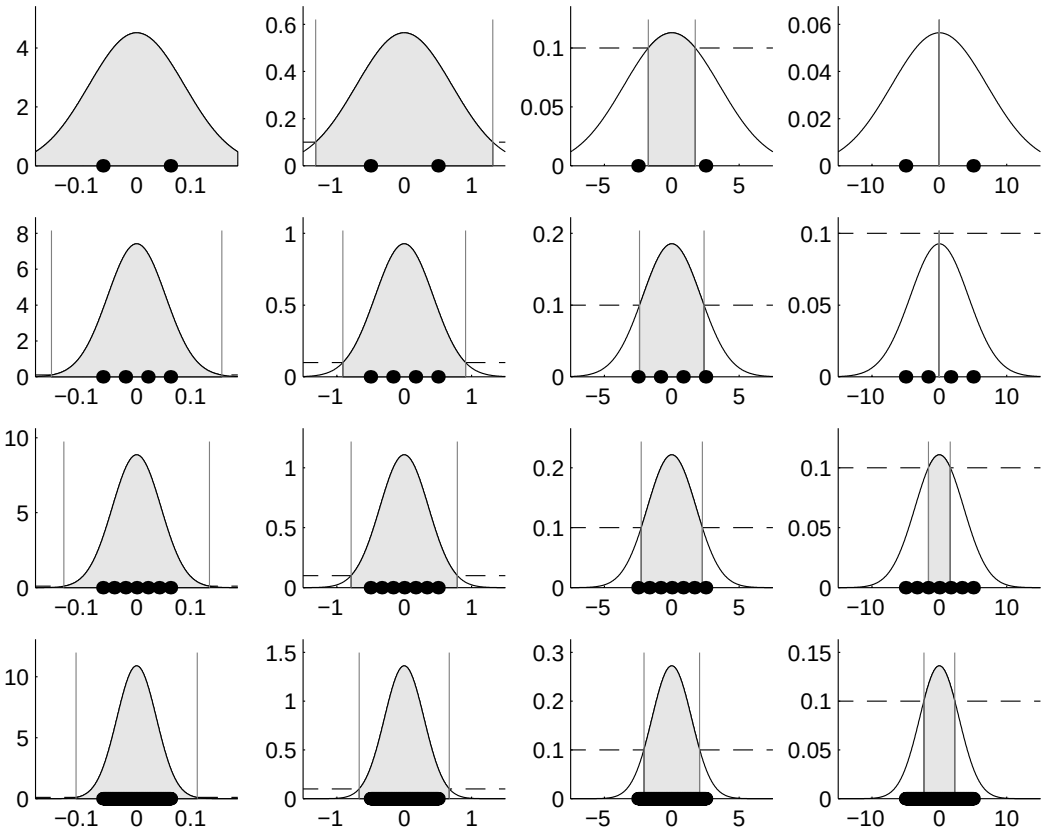

4
2
0
−0.1 0 0.1
0.6
0.4
0.2
0
−1 0 1
0.1
0.05
0
−5 0 5
0.06
0.04
0.02
0
−10 0 10
8
6
4
2
0
−0.1 0 0.1
1
0.5
0
−1 0 1
0.2
0.1
0
−5 0 5
0.1
0.05
0
−10 0 10
10
5
0
−0.1 0 0.1
1
0.5
0
−1 0 1
0.2
0.1
0
−5 0 5
0.1
0.05
0
−10 0 10
10
5
0
−0.1 0 0.1
1.5
1
0.5
0
−1 0 1
0.3
0.2
0.1
0
−5 0 5
0.15
0.1
0.05
0
−10 0 10

**Figure 4**

should always be judged the most certain instances of the concept, more certain than any novel stimulus. (Participants' judgments in all the experiments of Chapters 3-5 confirm this.) This may change if we expect that some of our observations could be mislabeled, but human learners naturally distinguish uncertainty in the observations from uncertainty about the extension of the concept, and we would like our theory to do the same. The Bayesian framework of this thesis does separate these two sources of ambiguity – the presentation up through Chapter 5 covers the noiseless case, while observation noise is considered in Appendix C. The novelty detection approach does not.

All distributional approaches have several other serious defects, resulting from their treatment of concepts as probability distributions. They have no way to account for dissociations between the frequency with which category members occur and their degree of membership in the category, observed by Rosch, Simpson & Miller (1976). They have no way to incorporate examples – either positive or negative – that are not randomly sampled from the concept. This kind of information occurs all the time: after you show me a few examples of healthy levels, I might then ask, "OK, how about this guy, are his levels healthy?", and I would like to be able to learn from your answer, even though it was not a random sample from either the healthy or unhealthy samples.

In summary, virtually all of the well-known similarity-based models of classification learning in the psychological literature can be interpreted either as direct discrimination learning algorithms or as distributional approaches implementing some version of Equation A.1 (Ashby & Leola-Reese, 1995). The direct approaches include all neural networks trained with supervised learning methods (Gluck & Bower, 1988; Shanks, 1988; Kruschke, 1992; Shanks & Gluck, 1994). The distributional approaches include prototype models (*e.g.* Fried & Holyoak, 1984), exemplar models (*e.g.* Nosofsky, 1986), and neural networks trained with unsupervised methods (*e.g.* Gluck & Myers, 1993). Anderson's (1991) "rational model" of categorization represents the same distributional information in a slightly different form, the joint statistics $p(\mathbf{v}, C)$ of features $\mathbf{v}$ and class labels $C$ (from which $p(\mathbf{v}|C)$, $p(\mathbf{v}| \sim C)$, and

$p(C)$ can be computed easily). All of these models, while often very successful for modeling classification learning tasks, are not capable of explaining how people can generalize concepts from just positive examples, for the reasons outlined in this appendix. Of course, this is not to say that these models couldn't be adapted somehow to our purposes, only that the adaptation would have to be fairly substantial.

It is instructive to compare these two standard approaches to classification and discrimination learning with the Bayesian framework of this thesis. In its basic representation of conceptual knowledge, the Bayesian framework is similar to direct discriminative approaches. That is, the Bayesian learner's basic hypotheses are rules for picking out the members of a concept, very much like the hard or soft classification rules that most direct discriminative approaches extract. The difference is that only the Bayesian concept learner is capable of scoring rules based on strictly positive evidence. Here, in its use of probabilistic assumptions to score hypotheses, the Bayesian framework is more similar to distributional approaches to classification. Both are capable of learning a model of a concept from only positive examples, based on maximizing the likelihood of the data given the model. The difference is that distributional approaches – because their model is a probability distribution for the positive instances – also require an analogous model of the negative examples in order to classify new objects via Equation A.1. In contrast, the Bayesian framework – because its model consists of rules that pick out the concept's instances – is capable of directly generalizing from only positive examples. Thus, it is the combination of rule-based hypotheses with a probabilistic generative model for scoring those hypotheses that distinguishes the Bayesian approach from standard models of discrimination and classification learning, and makes it appropriate for modeling concept learning from positive evidence only.

# Appendix B

# Derivation of generalization functions for continuous feature spaces

This appendix presents the detailed derivation of the generalization functions for continuous feature spaces, assuming a hypothesis space of all possible rectangular regions. All symbols not defined here are defined just as in Chapter 3. Recall that the generalization function $p(y \in C|X)$ is determined by integrating the predictions of all hypotheses, weighted by their posterior probabilities $p(h|X)$:

$$p(y \in C|X) = \int_{h \in \mathcal{H}} p(y \in C|h)\, p(h|X)\, dh, \tag{B.1}$$

where from Bayes' theorem $p(h|X) \propto p(X|h)p(h)$ (normalized such that $\int_{h \in \mathcal{H}} p(h|X)dh = 1$), and $p(y \in C|h) = 1$ if $y \in h$ and 0 otherwise. In order to compute the probability of generalization, we will use the equivalent form (equivalent to Equation 2.5 with sums replaced by integrals)

$$p(y \in C|X) = \frac{\int_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n dh}{\int_{h \in \mathcal{H}_X} p(h)/|h|^n dh}. \tag{B.2}$$

Recall the notation $\mathcal{H}_X$ and $\mathcal{H}_{X,y}$ for the subsets of hypotheses in $\mathcal{H}$ that contain $X$ and $X \cup \{y\}$ respectively. Each hypothesis $h$ in $\mathcal{H}$ can be indexed by four numbers $h(l_1, l_2, s_1, s_2)$, with $(l_1, l_2)$ denoting the location of its maximal value on each dimension (*i.e.* its upper-righthand corner) and $(s_1, s_2)$ denoting its size along each dimension.

## B.1 Uninformative prior

Under the uninformative prior (Equation 3.8), the generalization function has a simple closed-form expression valid for all $n \geq 2$. To see this, we first consider how to compute the denominator of Equation B.2, $p(X) = \int_{h \in \mathcal{H}_X} p(h)/|h|^n dh$. Once we have done this, the numerator $p(y \in C, X)$ is practically identical.

Because both $p(h)$ and $|h|$ are separable in the two dimensions of our feature space, we can just treat the one-dimensional version of this integral and then multiply the two one-dimensional integrals together to get the two-dimensional integral we want. (By the same argument, we could compute the generalization function for rectangle hypotheses in a space of any dimension $m$ by multiplying $m$ one-dimensional terms together.) In one dimension, we index each hypothesis by two numbers $l$ and $s$ defined as above. Because the problem as formulated is translation invariant, we can assume an arbitrary maximal value for the examples without loss of generality. We choose this maximum to be 0; thus all the examples have values less than or equal to 0 and all consistent hypotheses have $l \geq 0$. Finally we define $r$ to be the range spanned by the examples, *i.e.* the difference between the maximal example value (0) and the minimal example value. Then we have

$$
\begin{aligned}
p(X) &= \int_{h \in \mathcal{H}_X} \frac{p(h)}{|h|^n} dh && \text{(B.3)} \\
&= \int_r^\infty \int_0^{s-r} \frac{p(s)}{s^n} dl ds && \text{(B.4)} \\
&= \int_r^\infty \int_0^{s-r} \frac{1}{s^{n+1}} dl ds && \text{(B.5)} \\
&= \int_r^\infty \frac{s-r}{s^{n+1}} ds && \text{(B.6)}
\end{aligned}
$$

$$= \frac{-1}{(n-1)s^{n-1}}\Big|_r^\infty + \frac{r}{ns^n}\Big|_r^\infty \quad \text{(B.7)}$$

$$= \frac{1}{(n-1)r^{n-1}} - \frac{1}{nr^{n-1}} \quad \text{(B.8)}$$

$$= \frac{1}{n(n-1)r^{n-1}}. \quad \text{(B.9)}$$

Let $d$ denote the distance from the new stimulus $y$ to the closest observed example. We can assume that $y$ is outside the range spanned by the examples (otherwise the probability of generalization is simply 1), and without loss of generality, we can assume that $y$ has a positive value. Then to compute the numerator of Equation B.2, we can simply replace $r$ with $r + d$ in the limits of integration in the previous sequence of equations, yielding

$$p(y \in C, X) = \int_{h \in \mathcal{H}_{X,y}} \frac{p(h)}{|h|^n} dh \quad \text{(B.10)}$$

$$= \int_{r+d}^{\infty} \int_0^{s-(r+d)} \frac{p(s)}{s^n} dl ds \quad \text{(B.11)}$$

$$= \frac{1}{n(n-1)(r+d)^{n-1}}. \quad \text{(B.12)}$$

Putting numerator and denominator together, we obtain

$$p(y \in C|X) = \frac{\int_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n dh}{\int_{h \in \mathcal{H}_X} p(h)/|h|^n dh} \quad \text{(B.13)}$$

$$= \frac{n(n-1)r^{n-1}}{n(n-1)(r+d)^{n-1}} \quad \text{(B.14)}$$

$$= \frac{r^{n-1}}{(r+d)^{n-1}} \quad \text{(B.15)}$$

$$= \frac{1}{(1+d/r)^{n-1}}. \quad \text{(B.16)}$$

$$\quad \text{(B.17)}$$

Then combining the probability of generalization for each of the two dimensions of feature space yields the final answer reported in Chapter 3:

$$p_0(y \in C|X) = \left[\frac{1}{(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)}\right]^{n-1}. \quad \text{(B.18)}$$

Recall that the subscript "0" denotes the fact that using the uninformative prior corresponds to a gamma prior with $\alpha = 0$, and that $\tilde{d}_i$ is defined to be 0 if $y$ falls inside the range of values spanned by $X$ along dimension $i$, and otherwise is just $d$ as defined above, *i.e.* the distance from $y$ to the nearest example in $X$ along dimension $i$.

## B.2 Informative priors

Under the exponential prior (Equation 3.9), $p(y \in C|X)$ has no simple closed-form expression valid for all $n$. The same is true for the Erlang prior (Equation 3.10). However, we can derive an exact upper bound and an approximate lower bound on the generalization function using these priors, or any prior from the gamma family of densities (Equation 3.11),

$$p(s) \propto s^{\alpha-1} \exp\{-s/\sigma\}.$$

Adopting the same conventions and following the same reasoning as in the previous section, we can write the one-dimensional generalization function for any gamma prior as

$$\begin{aligned}
p(y \in C|X) &= \frac{\int_{h \in \mathcal{H}_{X,y}} p(h)/|h|^n dh}{\int_{h \in \mathcal{H}_X} p(h)/|h|^n dh} && \text{(B.19)} \\
&= \frac{\int_{r+d}^{\infty} \int_0^{s-(r+d)} \frac{p(s)}{s^n} dl ds}{\int_r^{\infty} \int_0^{s-r} \frac{p(s)}{s^n} dl ds} && \text{(B.20)} \\
&= \frac{\int_{r+d}^{\infty} \int_0^{s-(r+d)} \frac{1}{s^{n+1-\alpha}} \exp\{-s/\sigma\} dl ds}{\int_r^{\infty} \int_0^{s-r} \frac{1}{s^{n+1-\alpha}} \exp\{-s/\sigma\} dl ds} && \text{(B.21)} \\
&= \frac{\int_{r+d}^{\infty} \frac{s-(r+d)}{s^{n+1-\alpha}} \exp\{-s/\sigma\} ds}{\int_r^{\infty} \frac{s-r}{s^{n+1-\alpha}} \exp\{-s/\sigma\} ds}. && \text{(B.22)}
\end{aligned}$$

We then make the first of two changes of variables, substituting $t = s - (r+d)$ in the numerator and $t = s - r$ in the denominator. This gives

$$p(y \in C|X) = \frac{\int_0^{\infty} \frac{t}{(r+d)^{n+1-\alpha}(1+t/(r+d))^{n+1-\alpha}} \exp\{-(t+(r+d))/\sigma\} dt}{\int_0^{\infty} \frac{t}{r^{n+1-\alpha}(1+t/r)^{n+1-\alpha}} \exp\{-(t+r)/\sigma\} dt} \quad \text{(B.23)}$$

$$= \frac{\exp\{-d/\sigma\}}{(1+\frac{d}{r})^{n+1-\alpha}} \frac{\int_0^\infty \frac{t}{(1+t/(r+d))^{n+1-\alpha}} \exp\{-t/\sigma\} dt}{\int_0^\infty \frac{t}{(1+t/r)^{n+1-\alpha}} \exp\{-t/\sigma\} dt}. \quad \text{(B.24)}$$

A second change of variables, $u = t/(r+d)$ in the numerator and $u = t/r$ in the denominator, yields

$$p(y \in C|X) = \frac{\exp\{-d/\sigma\}}{(1+\frac{d}{r})^{n+1-\alpha}} \left(1+\frac{d}{r}\right)^2 \frac{\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-u(r+d)/\sigma\} du}{\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-ur/\sigma\} du} \quad \text{(B.25)}$$

$$= \frac{\exp\{-d/\sigma\}}{(1+\frac{d}{r})^{n-1-\alpha}} \left[\frac{\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-u(r+d)/\sigma\} du}{\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-ur/\sigma\} du}\right]. \quad \text{(B.26)}$$

Now, the bracketed term in Equation B.26 is always less than 1 (because the integrands of both numerator and denominator are always positive and the integrand of the numerator is strictly less than the integrand of the denominator). Thus we obtain the upper bound on generalization,

$$p(y \in C|X) \leq \frac{\exp\{-d/\sigma\}}{(1+\frac{d}{r})^{n-1-\alpha}}. \quad \text{(B.27)}$$

We can obtain an approximate lower bound by treating $u/(1+u)^{n+1-\alpha}$ as a windowing function on $\exp\{-ur/\sigma\}$ (in the denominator) or $\exp\{-u(r+d)/\sigma\}$ (in the numerator), and approximating it by a step function with value $c$ for $0 \leq u < \epsilon$ and value 0 for $\epsilon \leq u \leq \infty$, for some $c$ and $\epsilon$ that depend on $n$ and $\alpha$. The resulting approximation of the denominator,

$$\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-ur/\sigma\} du \approx c \int_0^\epsilon \exp\{-ur/\sigma\} du \quad \text{(B.28)}$$

$$= \frac{c\sigma}{r}\left[1 - \exp\{-\epsilon r/\sigma\}\right] \quad \text{(B.29)}$$

isn't a particularly good one, but the ratio of the corresponding approximation in the numerator to the above expression is, except for small $n$, a fairly good approximation to the quotient:

$$\frac{\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-u(r+d)/\sigma\} du}{\int_0^\infty \frac{u}{(1+u)^{n+1-\alpha}} \exp\{-ur/\sigma\} du} \approx \frac{1}{1+d/r}\left[\frac{1-\exp\{-\epsilon(r+d)/\sigma\}}{1-\exp\{-\epsilon r/\sigma\}}\right]. \quad \text{(B.30)}$$

Plugging the approximation in Equation B.30 into Equation B.26, we obtain another approximation for the generalization function:

$$p(y \in C|X) \approx \frac{\exp\{-d/\sigma\}}{(1+\frac{d}{r})^{n-\alpha}} \left[\frac{1-\exp\{-\epsilon(r+d)/\sigma\}}{1-\exp\{-\epsilon r/\sigma\}}\right]. \tag{B.31}$$

The bracketed term in Equation B.31 is always *greater* than 1. Hence this gives us an approximate lower bound for the generalization function,

$$p(y \in C|X) \gtrsim \frac{\exp\{-d/\sigma\}}{(1+\frac{d}{r})^{n-\alpha}}, \tag{B.32}$$

which is valid to the extent that the approximation in Equation B.30 is valid.

Conveniently, the two bounds in Equations B.27 and B.32 have the same functional form and differ only by one power of the denominator. Putting these bounds together and returing to two dimensions yields the expression cited in Chapter 3, (Equation 3.17)

$$p(y \in C|X) \approx \frac{\exp\{-(\tilde{d}_1/\sigma_1 + \tilde{d}_2/\sigma_2)\}}{\left[(1+\tilde{d}_1/r_1)(1+\tilde{d}_2/r_2)\right]^{n-\lambda}}, \tag{B.33}$$

which yields an upper bound for $\lambda = \alpha - 1$ and an approximate lower bound for $\lambda = \alpha$. In practice, I have found that the approximate lower bound (with $\lambda = \alpha$) provides a good approximation to the probability of generalization over a wide range of values for $n$ and $r_i$. For example, under the exponential prior ($\alpha = 1$), this approximation holds to within $\sim 10\%$ error, and usually much less, except for very small values of $n$ (e.g. $< 3$) and $r_i$ (e.g. $< \sigma_i/10$).

# Appendix C

# More complex inferences in concept learning (or, inferences based on hidden variables)

There are many complex inferences involved in real concept learning that I do not touch on in the body of this thesis. Examples include decisions about whether a particular example might be an outlier – an improperly labled example – and should be ignored in deciding how to generalize (Figure 1); decisions about whether two clusters of examples might correspond to two distinct extensions of a single concept label (as is the case with polysemy or homophony in language) (Figure 2); and decisions about which features from a large basis set might be relevant in defining the concept, when prior knowledge does not uniquely determine the relevant features (Figure 3). The outcomes of these decisions will have a significant impact on how we generalize a concept, particularly when we are learning from only a few positive examples. In this appendix, I sketch out how the Bayesian framework can be extended to handle these more subtle inferences in concept learning. For simplicity, I will focus on hypothesis spaces corresponding to rectangular regions in a continuous feature space (a la Chapter 3), although a similar analysis can be developed for other kinds of hypothesis spaces.

Inspiration for extending the Bayesian framework comes from, of all places, be-

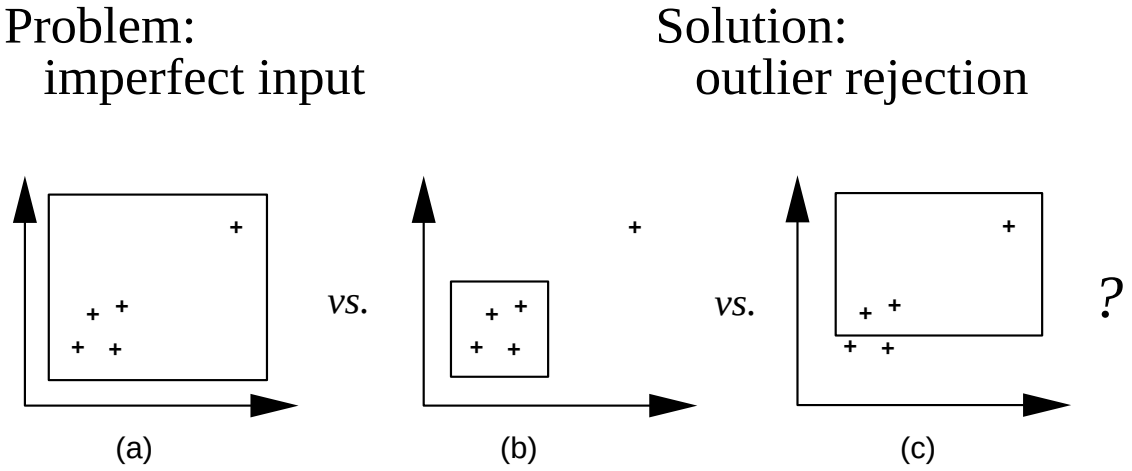



**Figure 1**

haviorist theories of concept leaning. Behaviorists thought of concept learning as a process of "mediated generalization" (Wasserman, 1995). That is, a concept is some kind of hidden variable in the organism's internal state that mediates between stimuli and responses. This idea has largely fallen by the wayside with the rise of cognitive science, but like many behaviorist notions, it contains an important grain of insight. "Hidden variables" are one of the main features of modern Bayesian and connectionist (e.g. Hinton et al., 1995; MacKay, 1992) models, which inspired this thesis' approach to concept learning. Hidden variables are often introduced to simply an otherwise intractable computation and then integrated out to yield the desired result. Here is an example of the hidden variable approach in probabilistic modeling. Suppose that we want to know the conditional probability of one variable $A$ given a second variable $B$, $p(A|B)$, but that for some reason, computing $p(A|B)$ directly is impossible or impractical. If we can find a third variable $H$, such that $A$ and $B$ are conditionally independent given $H$, and such that we can easily compute $p(A|H)$ and $p(H|B)$, then we now have a way to compute $p(A|B)$ by introducing $H$ as a hidden variable and integrating it out:

$$p(A|B) = \sum_H p(A, H|B) \quad \text{(C.1)}$$

$$= \sum_H p(A|H, B)p(H|B) \quad \text{(C.2)}$$

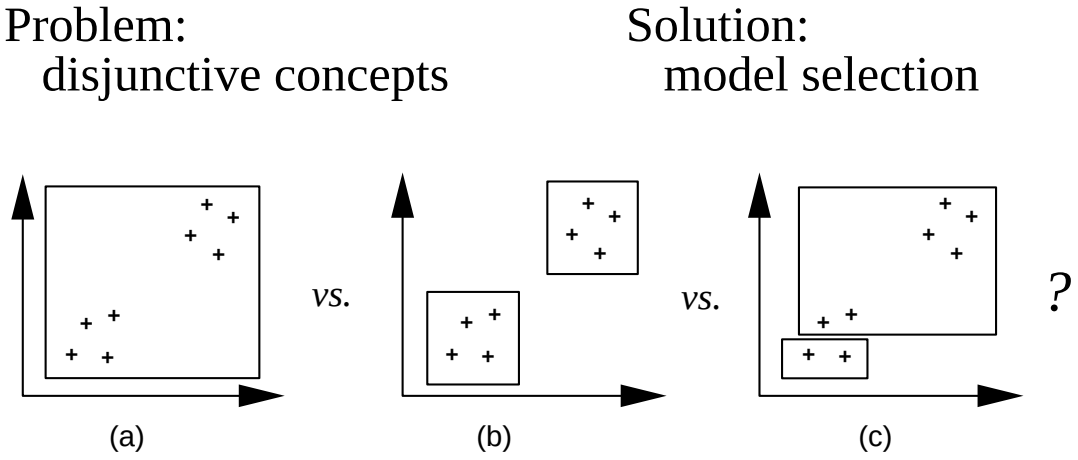



**Figure 2**

$$= \sum_H p(A|H)p(H|B) \tag{C.3}$$

Here the sum is taken over all possible values for $H$. The third equation follows from the second by the assumption that $A$ and $B$ are conditionally independent given $H$.

It turns out that the Bayesian analysis of concept learning conforms precisely to this abstract mathematical form. We replace $B$ by "$X = \{x_1, \ldots, x_n\}$", the event that we have observed these $n$ examples of a concept $C$, and we replace $A$ by "$y \in C$", the event that a new entity $y$ belongs to $C$. The hidden variable $H$ corresponds to the unknown extension of the concept. If the extension were known (*i.e.* given $H$), the examples $X$ and the event $y \in C$ become independent (because $y$ either belongs to the hypothetical extension $H$ or it doesn't, independently of any other observations). Thus $p(A|H, B) = p(A|H)$ becomes 1 if $y \in H$ and 0 otherwise. The probability $p(H|B)$ becomes the posterior probability of the hypothesis given the examples, which we saw how to calculate in Chapter 2. Making these substitutions into Equation C.3 produces exactly Equation 3 of Chapter 2, the fundamental equation of Bayesian generalization.

Once we have made the connection between Bayesian generalization and hidden variable modeling, there is nothing to stop us from introducing other hidden variables to compute more elaborate inferences. By introducing hidden variables denoting the possibility of outliers, disjunctive (*i.e.* polysemic or homophonic) concepts, and the

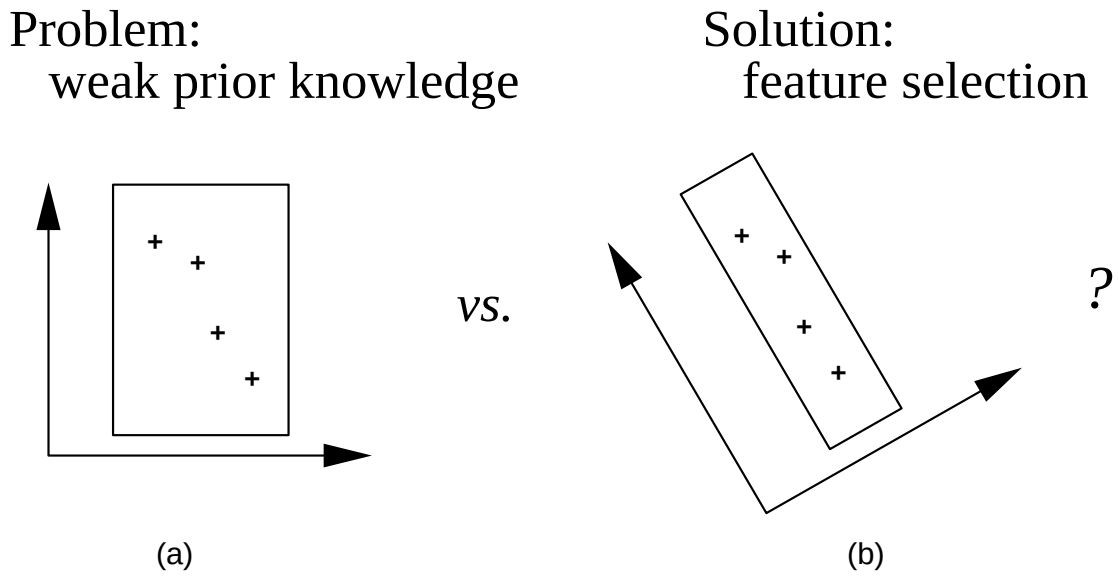



**Figure 3**

relevant feature axes, we can model these more sophisticated inferences that must be a part of concept learning in the real world. Figures 4 and 5 show the effects of hidden variables for outliers and disjunctive concepts on generalization gradients in a 1-dimensional version of the healthy levels task of Chapter 3, *i.e.* where the stimuli correspond to points along a line and the hypotheses correspond to intervals of that line. The outlier hidden variable ranges over all proper subsets of the examples, while the disjunction hidden variables ranges over all ways of partitioning the examples into two distinct subsets (drawn independently from two independent extensions). The left-hand column shows the resulting generalization gradients; the right-hand column shows the most likely states of the hidden variables, along with their probabilities. Figure 6 shows the effect of a hidden variable for relevant feature axes, which ranges over all possible orientations in a two-dimensional continuous feature space with rectangle hypotheses. [1] All three of these inferences depend on the same foundation that guided generalization throughout this thesis: the size principle and hypothesis averaging. In the rest of this appendix, I give a sketch of the mathematics behind these inferences. A full treatment will be the subject of a future paper.

[1] This is one possible way of modeling generalization in psychological spaces with "integral" dimensions. See Shepard (1987) for another proposal.

## C.1 Outlier rejection for imperfect input

Assume the examples $X$ are partitioned into two subsets, a set of *inliers* $X_{in}$ and a set of *outliers* $X_{out} = X \; X_{in}$. The inliers are actual random samples from the concept $C$ to be learned, while the outliers are produced by some other unknown process which has nothing to do with the concept to be learned. Thus, our generalization behavior should be based solely on the inferences we draw from the inliers; the outliers should be ignored. The problem is that we do not know which, if any, examples are actually outliers. All we have are the raw labeled examples, although intuitively, some examples seem more likely to be outliers than others. For instance, in Figure, the lone example in the upper-right corner seems much more likely than any of the other four examples to be an outlier. This section shows how a Bayesian learner can infer the likely outliers and then, based on this inference, decide the correct way to generalize the concept. For simplicity, I will work with a one-dimensional version of the axis-parallel rectangles task. The generalization to other hypothesis spaces is straightforward.

Let $w$ denote a particular partition of $X$ into $\{X_{in}, X_{out}\}$. If there are a total of $n$ examples observed, then $w$ can take on one of $2^n - 1$ possible states (assuming we are not going to reject every example as an outlier). If we knew the true value of $w$ – the actual identities of the inliers and outliers in our example set – then we would know how to calculate the generalization function:

$$p(y \in C|w, X) = p(y \in C|X_{in}). \tag{C.4}$$

That is, the generalization function given knowledge of $w$ is just the standard generalization function $p(y \in C|X)$ (derived for axis-parallel rectangles in Chapter 3 and Appendix B, Equation B.17), with $X$ replaced by $X_{in}$. However, because we don't know the true value of $w$, we introduce it as a hidden variable and then integrate it

out:

$$p(y \in C|X) = \sum_w p(y \in C, w|X) \quad \text{(C.5)}$$

$$= \sum_w p(y \in C|w, X)\ p(w|X) \quad \text{(C.6)}$$

$$= \sum_w p(y \in C|X_{in})\ p(w|X). \quad \text{(C.7)}$$

Equation C.7 tells us that the optimal way to generalize in the presence of unknown outliers is to average the simple generalization functions computed for each possible set of inliers $w$ weighted by $p(w|X)$, the posterior probability that $w$ is in fact the set of outliers given the observations $X$. We calculate $p(w|X)$ using Bayes' rule under the assumption of independently sampled examples:

$$p(w|X) \propto p(X|w)p(w) \quad \text{(C.8)}$$

$$\propto p(X_{in}|w)p(X_{out}|w)p(w). \quad \text{(C.9)}$$

Assuming an uninformative (scale-invariant) prior and inliers randomly sampled from a uniform distribution over the concept,

$$p(X_{in}|w) = \frac{1}{n_{in}(n_{in}-1)r_{in}^{n_{in}-1}}, \quad \text{(C.10)}$$

where $n_{in}$ denotes the number of inliers indexed by $w$ and $r_{in}$ denotes the range spanned by the inliers (*i.e.* the size of the smallest interval containing all inliers).[2] Equation C.10 follows from the derivation in Appendix B and is essentially the same

[2]This treatment brushes over one complication, namely that Equation C.10 is not defined for inlier sets containing only one example. One solution is to compute the generalization function using an informative prior, which would then be defined for the $n = 1$ case. An alternative approach was adopted in the computations behind Figures 4 and 5, in order to ensure an answer in the same algebraic form as the generalization functions obtained under an informative prior. We first assume a prior density on the size $s$ of the form $p(s) \propto 1/s^{1+\beta}$ for some small $\beta$ (which reduces to the uninformative prior in the limit $\beta \to 0$). We also assume that there is a "just noticeable difference" $\alpha$ within which stimuli are indistinguishable. This requires adding $\alpha$ to every range; $r$ becomes $r+\alpha$, $r_{in}$ becomes $r_{in}+\alpha$, and so on. The generalization function then becomes, instead of Equation B.17,

$$p(y \in C|X) = \frac{1}{[1 + d/(r+\alpha)]^{n+\beta-1}}. \quad \text{(C.11)}$$

as Equation B.9 with $r$ replaced by $r_{in}$. If we assume that there is some small probability $\epsilon$ of seeing an outlier on any particular learning trial and that outliers are drawn independently from a uniform density over a region of size $L$, then we can combine the last two terms of Equation C.9 into one,

$$p(X_{out}|w)p(w) \propto (\epsilon/L)^{n_{out}}, \tag{C.13}$$

in which $n_{out} = n - n_{in}$ is the number of outliers indexed by $w$. The ratio $\epsilon/L$ affects the precise predictions of the model but not the qualitative behavior. The posterior probability of any particular division into inliers and outliers then becomes

$$p(w|X) \propto \left(\frac{1}{r_{in}}\right)^{n_{in}-1} \left(\frac{\epsilon}{L}\right)^{n_{out}}. \tag{C.14}$$

Notice how $p(w|X)$ balances two intuitively relevant opposing constraints to find the optimal size and composition of the inlier set. As the number of outliers $n_{out}$ increases (and the number of inliers $n_{in}$ decreases), the inlier range $r_{in}$ necessarily decreases or stays the same. Increasing $n_{out}$ acts to decrease $p(w|X)$ at the same time as decreasing $r_{in}$ tends to *increase* $p(w|X)$. In other words, the decision to reject any particular example as an outlier is always penalized by a constant factor, but it may lead to a net increase in probability if it dramatically reduces the size $r_{in}$ of the inlier set. This explains why the example in the upper-right corner of Figure 1 seems more likely to be an outlier than any of the other examples, because rejecting it would lead to a huge decrease in the size of the inlier set (Figure 1b), while rejecting, say, two of the others would lead to only a small decrease in the size of the inlier set (Figure 1c). This procedure for rational outlier rejection is yet another case of the size principle emerging from Bayesian inference to provide a principled form of Ockham's razor.

---

Equation C.10 becomes

$$p(X_{in}|w) = \frac{1}{(n_{in}+\beta)(n_{in}+\beta-1)(r_{in}+\alpha)^{n_{in}+\beta-1}}. \tag{C.12}$$

Figure 4 was generated by setting $\alpha$ to roughly 1% of the visible range, and $\beta$ to 0.2. Qualitatively similar results are obtained for other values of $\alpha$ and $\beta$, as long as they are kept reasonably small but bounded away from zero.

Figure 4 shows the generalization functions for four different distributions of examples in a one-dimensional space with rectangle hypotheses. Observe that generalization is no longer constant across the entire range of examples but reflects the density of examples in an intuitively reasonable way. The extent to which one observation is separated from the main cluster, and the density of the cluster, determine the probability with which that observation will be classified as an outlier and hence receive less than a 100% probability of generalization. Also depicted in Figure 4 are the 15 most probable inlier sets (*i.e.* values of $w$), along with their probabilities. Note that in each case, $p(w|X)$ essentially goes to zero beyond the four or five most probable values of $w$, and that the most probable inlier sets are all quite similar to each other. Thus the average over all possible states of $w$ could probably be well approximated by rapid Gibbs sampling (Geman & Geman, 1984), avoiding the need for any intractable sums in the generalization procedure.

## C.2 Model selection for disjunctive concepts

Assume the examples $X$ are partitioned into two subsets, a set $X_1$ drawn from one extension in the hypothesis space and a second set $X_2 = X \setminus X_1$ drawn from an independently chosen extension in the same hypothesis space. Both $X_1$ and $X_2$ are assumed to be random samples from their respective extensions. This situation might occur when a learning a word with two distinct senses (polysemy), or when learning two different words that are phonologically identical (homophony), and hence might seem like a single word with two distinct extensions until the homophony is recognized. As in the previous section, we would like to infer whether and how the examples were drawn from two distinct and independent extensions, in order to generalize appropriately. Intuitively, some splits seem more reasonable than others – compare Figure 2b with 2c – and, as before, the relative ranges spanned by the resulting groups seems to be an important factor in determining this preference. Again, for simplicity I will work with a one-dimensional version of the axis-parallel rectangles task, but the generalization to other hypothesis spaces is straightforward.

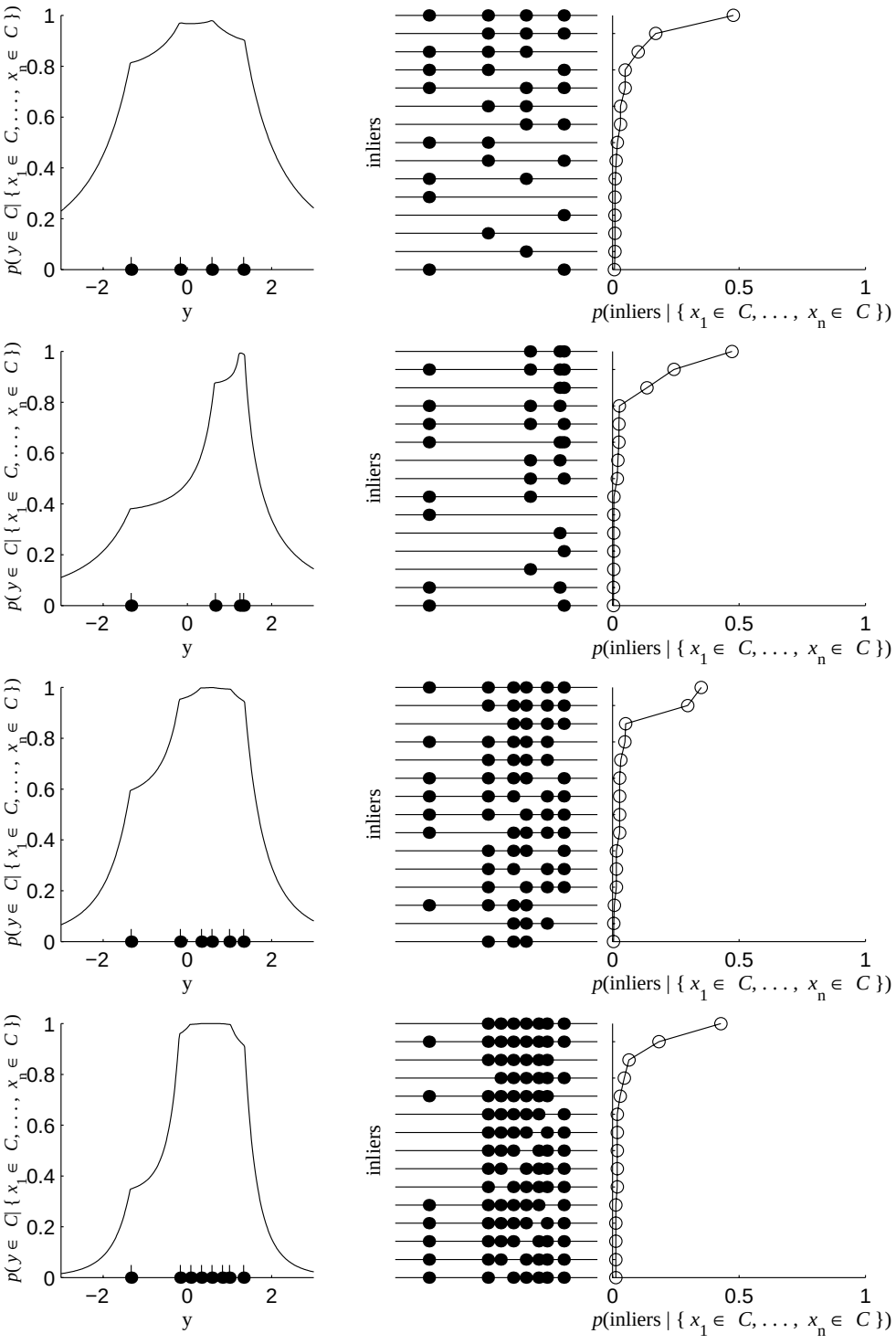



**Figure 4**

Let $z$ denote a particular partition of $X$ into $\{X_1, X_2\}$. If there are a total of $n$ examples observed, then $z$ can take on one of $2^n - 1$ possible states. [3] If we knew the true value of $z$ – how the examples were actually grouped into $X_1$ and $X_2$ – then we would know how to calculate the generalization function. Assuming that the two extensions are independent (*i.e.* independent samples from $p(h)$) and that the examples drawn from each extension are also drawn independently, it can be shown that

$$p(y \in C|z, X) = p(y \in C|X_1) \; + \; p(y \in C|X_2) \; - \; p(y \in C|X_1) \; p(y \in C|X_2). \quad \text{(C.15)}$$

That is, the generalization function given knowledge of $z$ is just the disjunctive combination of the two standard generalization functions $p(y \in C|X_1)$ and $p(y \in C|X_2)$ for each group separately (derived for axis-parallel rectangles in Chapter 3 and Appendix B). However, because we don't know the true value of $z$, we introduce it as a hidden variable and then integrate it out:

$$\begin{aligned} p(y \in C|X) &= \sum_z \; p(y \in C, z|X) && \text{(C.16)} \\ &= \sum_z \; p(y \in C|z, X) \; p(z|X) && \text{(C.17)} \\ &= \sum_z \; [p(\bullet|X_1) + p(\bullet|X_2) - p(\bullet|X_1)p(\bullet|X_2)] \; p(z|X), && \text{(C.18)} \end{aligned}$$

where $\bullet$ stands for "$y \in C$". This amounts to averaging the disjunctive combinations of generalization functions resulting from all possible splits of the examples in $X$, weighted by the posterior probability $p(z|X)$ of each split $z$. We calculate $p(z|X)$ from Bayes' rule (continuing under the assumption of independently sampled examples):

$$\begin{aligned} p(z|X) &\propto p(X|z)p(z) && \text{(C.19)} \\ &\propto p(X_1|z)p(X_2|z)p(z). && \text{(C.20)} \end{aligned}$$

[3] Because there is no difference between $X_1$ being empty and $X_2$ being empty, we assume without loss of generality that $n_2 > 0$.

Assuming an uninformative (scale-invariant) prior and examples randomly sampled from a uniform distribution over the appropriate extension,

$$p(X_1|z) = \frac{1}{n_1(n_1-1)r_1^{n_1-1}}, \tag{C.21}$$

where $n_1$ denotes the number of examples in $X_1$ as indexed by $z$, and $r_1$ denotes the range spanned by those examples (*i.e.* the size of the smallest interval containing them).[4] The analogous expression gives $p(X_2|z)$:

$$p(X_2|z) = \frac{1}{n_2(n_2-1)r_2^{n_2-1}}, \tag{C.22}$$

with $n_2 = n - n_1$ and $r_2$ defined accordingly. Finally, we set $p(z)$ by assuming that with probability $\kappa$ the examples are drawn from a single extension (*i.e.* $p(z) = \kappa$, for $n_1 = 0$), while with probability $1 - \kappa$ they are drawn from two distinct extensions in some arbitrary manner (*i.e.* $p(z) = (1 - \kappa)$, for $n_1 > 0$).

The result of combining these three terms into the posterior $p(z|X)$ is again to balance several opposing forces. Switching the assignment of examples from $X_1$ to $X_2$ decreases (or leaves unchanged) the range $r_1$ but *increases* (or leaves unchanged) the range $r_2$. The optimal decision about which group to assign a particular example to is thus determined by which group's range it will have a greater effect on. This accounts for why the split in Figure 2b appears more natural than the split in Figure 2c. In addition, while the special case of assigning all the examples to one group is the least favored on relative size grounds, it is favored on other grounds (dependent on $\kappa$ and $L$). This accounts for why a set of examples must be highly clustered – giving a strong relative size benefit as in Figure 2 – in order to be a plausible candidate for

[4] As in the previous section, this treatment ignores the fact that Equation C.21 is not defined when $X_1$ contains only one example. We deal with this complication using the same procedure that was described at the analogous point of the previous section, with analogous results that depend on $\alpha$ and $\beta$ parameters. Figure 5 uses exactly the same values for these parameters as does Figure 4, and again the results do not depend critically on these choices. A more serious problem is the case $n_1 = 0$, *i.e.* when all the examples are drawn from a single extension. A rigorous analysis should take into account the volume of the hypothesis space, which is impossible under the improper uninformative prior. A heuristic approximation used to generate Figure 5 is to set $p(X_1|z) = L$ for the case of $n_1 = 0$, where $L$ is the range of possible stimulus values visible in the figure.

having arisen from two independent extensions.

Figure 5 shows the generalization functions for four different distributions of examples in a one-dimensional space with rectangle hypotheses. As in Figure 4, generalization is no longer constant across the entire range of examples but reflects the density of examples in intuitively reasonable ways. The degree of bimodality in the generalization function reflects the degree to which the data provide support for the existence of two independent extensions, which in turn depends on several factors: the relative ranges spanned by the two clusters of data, the separation between clusters, and total number of data points observed. Also depicted are the 8 most probable splits into two independent sets of examples (*i.e.* the 8 most probable values of $z$), along with their probabilities. Again, the distributions of $p(z|X)$ are for the most part sharply peaked about their maximum values, making them candidates for efficient computation by stochastic simulation.

## C.3 Feature selection under weak prior knowledge

Suppose that our prior knowledge is weaker than we have previously assumed. We know that the concept to be learned corresponds to some rectangular region in a two-dimensional feature space, but we do not know the correct set of axes for this space, *i.e.* the correct orientation of the concept in feature space. Another case of weak prior knowledge might occur if we have a very high-dimensional space of potentially relevant features, and we know that the concept corresponds to a rectangle in some 2-dimensional (or $k$-dimensional) subspace of this high-dimensional feature space, but we don't know which subspace. Bayesian inference again comes to our aid, telling us which of these different hypothesis spaces are most likely to be the true hypothesis space that the observed examples were drawn from. Here I just the consider the case of a concept corresponding to a rectangular region of unknown orientation in a two-dimensional feature space (Figure 3).

Let $\theta \in [0, \pi/2]$ denote a particular orientation of the axes in feature space. If we knew the true value of $\theta$ then we would know how to calculate the generalization

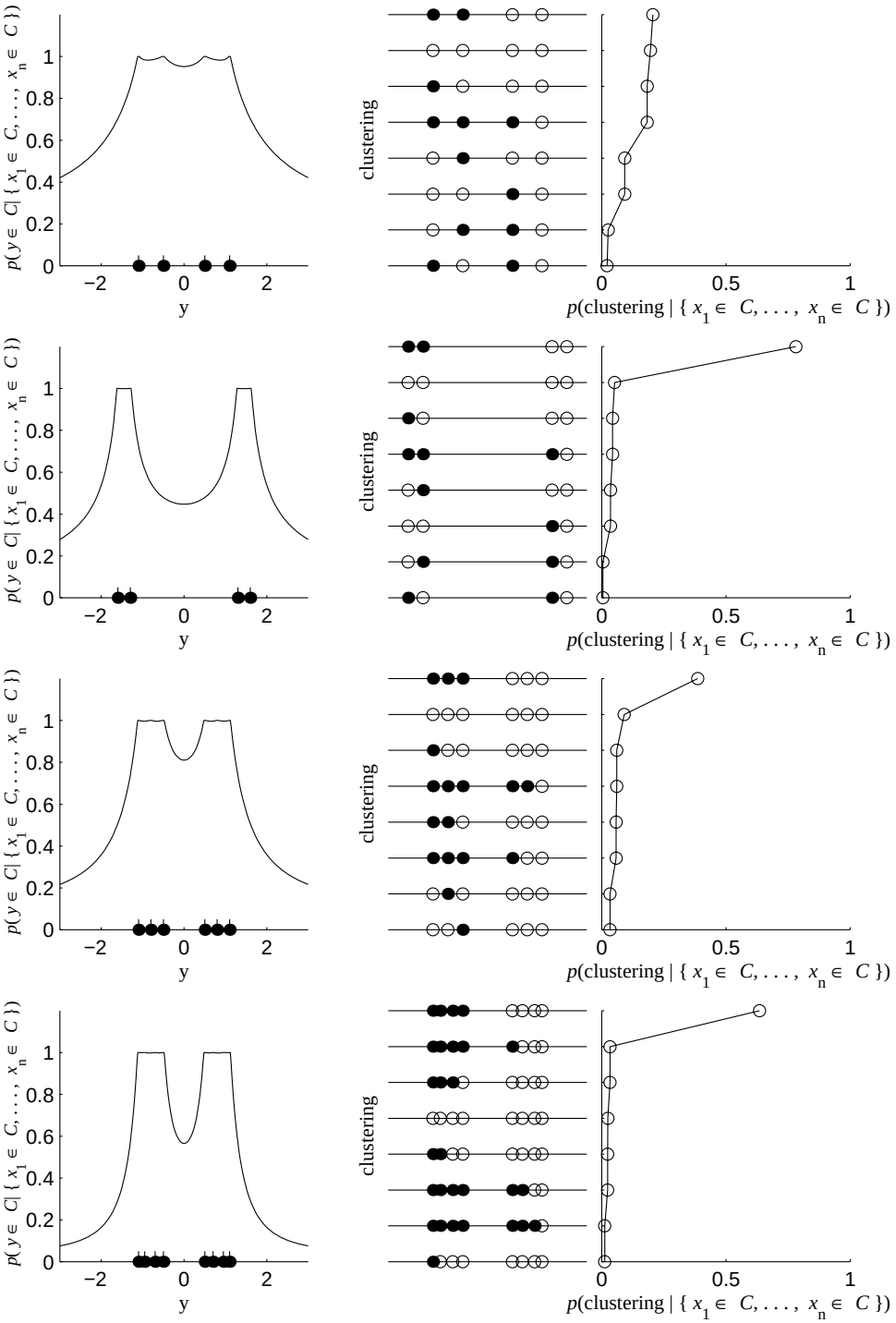



**Figure 5**

function $p(y \in C|\theta, X)$, using the standard machinery of Chapter 3 and Appendix B applied to the space of all rectangles parallel to the axes at orientation $\theta$. However, because we don't know the true value of $w$, we introduce it as a hidden variable and then integrate it out:

$$p(y \in C|X) = \sum_{\theta} p(y \in C, \theta|X) \quad \text{(C.23)}$$

$$= \sum_{\theta} p(y \in C|\theta, X)\ p(\theta|X). \quad \text{(C.24)}$$

As before, we simply average the generalization functions at each setting of the hidden variable $\theta$ weighted by the posterior probability $p(\theta|X)$. From Bayes' rule,

$$p(\theta|X) \propto p(X|\theta)p(\theta). \quad \text{(C.25)}$$

The prior on orientation $p(\theta)$ can embody any *a priori* knowledge we do have favoring some particular axes for feature space. Here we assume no such knowledge, *i.e.* $p(\theta) = 2/\pi$ for $\theta \in [0, \pi/2]$. Assuming an uninformative (scale-invariant) prior on rectangle size and location, and given $n$ examples randomly sampled from a uniform distribution over the concept, the likelihood function for orientation $\theta$ is given by:

$$p(X|\theta) \propto \frac{1}{(r_\theta r_{\theta+\pi/2})^{n-1}}, \quad \text{(C.26)}$$

where $r_\theta$ denotes the range spanned by the examples when projected onto a line at orientation $\theta$. [5] This likelihood favors orientations onto which the data project as *small* a range as possible, which accounts for the preference of some axes (*e.g.* Figure 3b) over others (*e.g.* Figure 3a).

Figure 6 shows the generalization functions for four different distributions of examples in a two-dimensional space. Observe that the blocky and sharp-cornered generalization contours of Chapter 3 have been replaced by smoother contours, as a consequence of averaging over all possible rectangle orientations. Also depicted is the

[5] We must be careful in that unless we have observed more examples than there are dimensions in the feature space, there will always be some value of $\theta$ that makes $p(X|\theta)$ infinite.

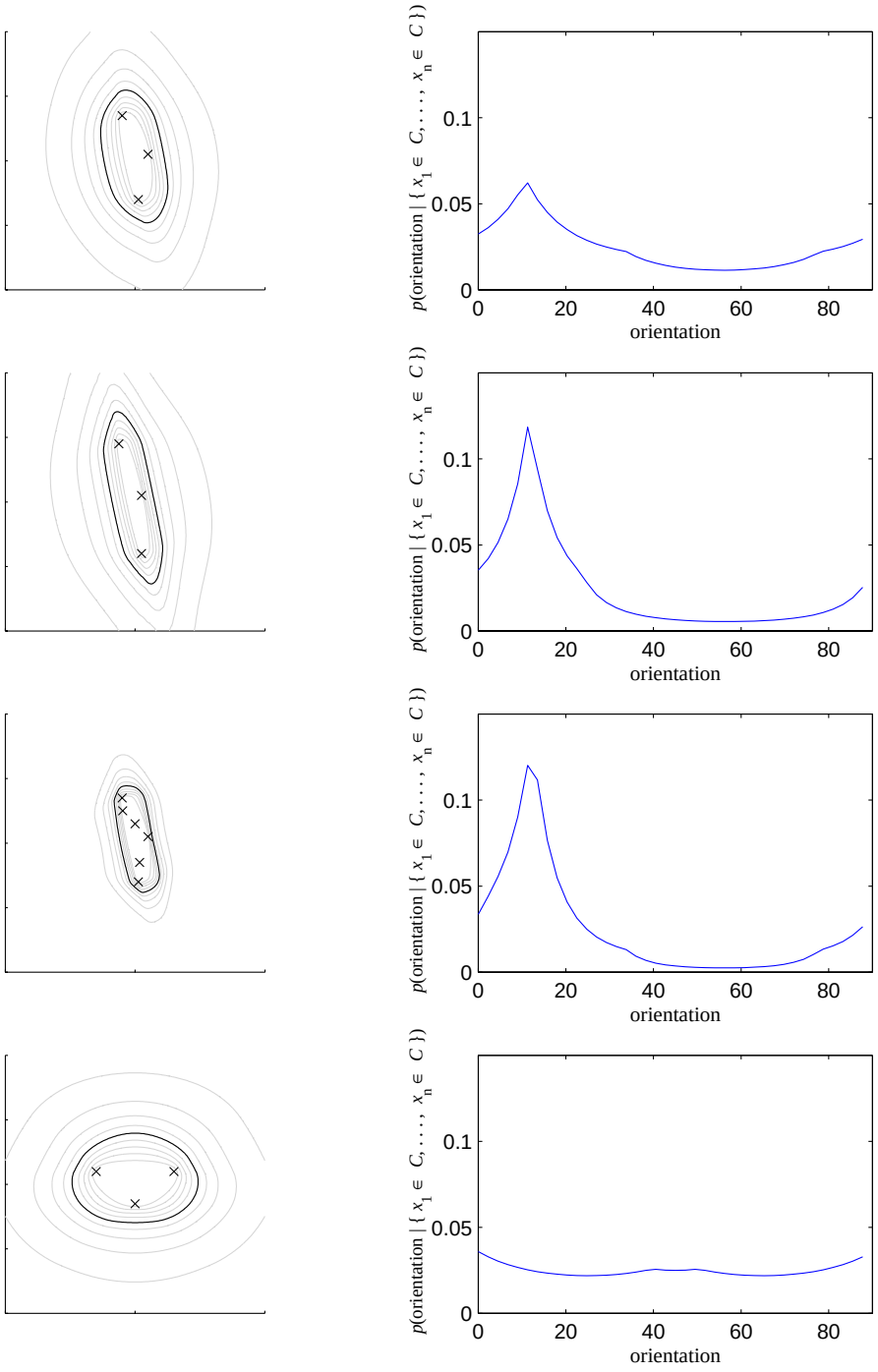



**Figure 6**

posterior probability $p(\theta|X)$ of each axis orientation. Notice how the orientation at which the data are most tightly clustered receives the highest probability assignment, and that the resulting gradients of generalization are correspondingly elongated or compressed along the data's principal axes. The extent to which the generalization gradients deviate from circularity is a function of how much evidence the data provide for a preferred set of axes, *i.e.* how tightly the examples are along any one direction as well as how many examples have been observed.

# Bibliography


Aha, D. and Goldstone, R. (1992). Learning attribute relevance in context in instance-based learning algorithms. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 14. L. Erlbaum.

Aha, D. W., editor (1997). *Lazy learning*. Kluwer.

Allen, S. W. and Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120(1):3–19.

Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13:263–308.

Ashby, F. G. (1992). *Multidimensional models of perception and cognition*. Erlbaum.

Ashby, F. G. and Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39:216–233.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105(4):442–481.

Atran, S. (1998). Folk biology and the anthropology of science: Cultural universals and cultural particulars. *Behavioral and Brain Sciences*, 21(4):547–609.

Attneave, F. (1959). *Applications of information theory to psychology; A summary of basic concepts, methods, and results.* Holt.

Bacon, F. (1620/1960). *The new organon, and related writings.* Bobbs-Merrill.

Barlow, H. (1985). Cerebral cortex as a model builder. In Rose, D. and Dobson, V. G., editors, *Models of the Visual Cortex*, pages 37–46. Wiley.

Barlow, H. (1996). Banishing the homunculus. In Knill, D. C. and Richards, W. A., editors, *Perception as Bayesian Inference*, pages 425–450. Cambridge University Press.

Barrow, H. G., Bolles, R. C., Garvey, T. D., Kremers, J. H., Lantz, K., Tenenbaum, J. M., and Wolf, H. C. (1977). Interactive aids for cartography and photo interpretation. In *Proceedings of the ARPA Image Understanding Workshop*, pages 111–127. Science Applications Inc.

Barsalou, L. (1985). Ideals, central tendencies, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4):629–654.

Barsalou, L. W., Huttenlocher, J., and Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, 36(3):203–272.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer.

Berwick, R. C. (1985). *The acquisition of syntactic knowledge.* MIT Press.

Binford, T. O. (1981). Inferring surfaces from images. *Artificial Intelligence*, 17:205–244.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford.

Bloom, P. (in press). Theories of word learning: Rationalist alternatives to associationism. In Bhatia, T. K. and Ritchie, W. C., editors, *Handbook of language acquisition.* Academic Press, New York.

Bloom, P. and Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Science*, 2:67–73.

Bobick, A. (1987). *Natural Object Categorization*. PhD thesis, Department of Brain and Cognitive Sciences, MIT.

Bower, G. H. and Trabasso, T. R. (1964). Concept identification. In Atkinson, R. C., editor, *Studies in mathematical psychology*, pages 32–94. Stanford University Press, Stanford, CA.

Brase, G. L., Cosmides, L., and Tooby, J. (1998). Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty. *Journal of Experimental Psychology: General*, 127(1):3–21.

Brent, M. R. and Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61:93–125.

Brown, R. (1958). How shall a thing be called? *Psychological Review*, 65:14–21.

Bruner, J. A., Goodnow, J. S., and Austin, G. J. (1956). *A study of thinking*. Wiley.

Callanan, M. A. (1989). Development of object categories and inclusion relations: Preschoolers' hypotheses about word meanings. *Developmental Psychology*, 25(2):207–216.

Callanan, M. A., Repp, A. M., McCarthy, M. G., and Latzke, M. A. (1994). Children's hypotheses about word meanings: Is there a basic level constraint? *Journal of Experimental Child Psychology*, 57:108–138.

Carey, S. (1978). The child as word learner. In Halle, M., Bresnan, J., and Miller, G. A., editors, *Linguistic theory and psychological reality*, pages 264–293. MIT Press, Cambridge, MA.

Carey, S. (1982). Semantic development: The state of the art. In Wanner, E. and Gleitman, L. R., editors, *Language Acquisition: The state of the art*, pages 347–389. Cambridge University Press, New York.

Carey, S. (1985a). *Conceptual change in childhood*. MIT Press.

Carey, S. (1985b). Constraints on semantic development. In Mehler, J. and Fox, R., editors, *Neonate cognition: Beyond the blooming, buzzing confusion*, pages 381–398. Erlbaum, Hillsdale, NJ.

Chomsky, N. (1986). *Language and problems of knowledge: The Managua lectures*. MIT Press.

Clark, E. V. (1973). What's in a word? on the child's acquisition of semantics in his first language. In Moore, T. E., editor, *Cognitive Development and the Acquisition of Language*. Academic Press.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In MacWhinney, B., editor, *The 20th Annual Carnegie Symposium on Cognition*. Erlbaum.

Coley, J. D., Medin, D. L., and Atran, S. (1998). Does rank have its privelege? inductive inferences within folkbiological taxonomies. *Cognition*, 64(1):73–112.

Corter, J. E. and Tversky, A. (1986). Extended similarity trees. *Psychometrika*, 51(3):429–451.

Cosmides, L. and Tooby, J. (1992). Cognitive adaptations for social exchange. In Barkow, J., Cosmides, L., and Tooby, J., editors, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press.

Cosmides, L. and Tooby, J. (1996). Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58:1–73.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13:21–27.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

DeBonet, J. S. and Viola, P. (1998). Structure driven image database retrieval. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Dietterich, T. G., Lathrop, R. H., and Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–72.

Dretske, F., editor (1981). *Knowledge and the Flow of Information*. MIT Press.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis*. Wiley.

Earman, J., editor (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press.

Elio, R. and Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory and Cognition*, 12(1):20–30.

Erickson, M. A. and Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127(2):107–140.

Estes, W. K. (1994). *Classification and cognition*. Oxford University Press.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, 41:145–170.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A., editor (1983). *The Modulariy of Mind: an Essay on Faculty Psychology*. MIT Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press.

Forster, M. R. The new science of simplicity. In Keuzenkamp, H., McAleer, M., and Zellner, A., editors, *Simplicity, Inference, and Econometric Modelling*. Cambridge University Press.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545.

Fried, L. S. and Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10:234–257.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. Wiley.

Garner, W. R. (1974). *The processing of information and structure*. Erlbaum.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemna. *Neural Computation*, 4(1):1–58.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25:407–454.

Gigerenzer, G. (1996). On narrow norms and vague heuristics. *Psychological Review*, 103(3):592–596.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102:684–704.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, 2(1):50–55.

Gluck, M. A. and Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3):227–247.

Gluck, M. A. and Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3(4):491–516.

Goldman, A. I. (1986). *Epistemology and Cognition*. Harvard University Press.

Goldstone, R. L. (1994a). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2):178–200.

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52(2):125–157.

Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.

Goodman, N. (1972). Seven strictures on similarity. In Goodman, N., editor, *Problems and projects*. The Bobbs-Merrill Co.

Gull, S. F. (1989). Developments in maximum entropy data analysis. In Skilling, J., editor, *Maximum Entropy and Bayesian Methods, Cambridge, 1988*, pages 53–71. Kluwer.

Haussler, D. (1988). Quantifying inductive bias: Ai learning algorithms and valiant's learning framework. *Artificial Intelligence*, 36:177–221.

Haussler, D., Kearns, M., and Schapire, R. E. (1994). Bounds on the sample complexity of bayesian learning using information theory and the vc-dimension. *Machine Learning*, 14:83–113.

Heit, E. (1997). Knowledge and concept learning. In Lamberts, K. and Shanks, D., editors, *Knowledge, concepts, and categories*, pages 7–42. MIT Press.

Heit, E. (1998). A bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., editors, *Rational models of cognition*. Oxford University Press.

Hinton, G., Dayan, P., Frey, B., and Neal, R. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161.

Hofstadter, D. (1995). *Fluid Concepts and Creative Analogies.* Basic Books.

Horwich, P., editor (1982). *Probability and Evidence.* Cambridge University Press.

Horwich, P. (1993). Wittgensteinian bayesianism. In French, P. A., T. E. Uehling, J., and Wettstein, H. K., editors, *Midwest Studies in Philosophy*, volume 18, pages 62–77. Notre Dame Press.

Hovland, C. I. (1952). A 'communication-analysis' of concept learning. *Psychological Review*, 59:461–472.

Howson, C. and Urbach, P. (1989). *Scientific reasoning: the Bayesian approach.* Open Court.

Hume, D. (1739/1978). *A treatise of human nature.* Clarendon Press.

Hunt, E. B. (1962). *Concept learning, an information processing problem.* Wiley.

Jaakola, T., Saul, L. K., and Jordan, M. I. (1996). Fast learning by bounding likelihoods in sigmoid type belief networks. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 528–534. MIT Press.

James, W. (1890/1981). *The principles of psychology.* Harvard University Press.

Japkowicz, N., Myers, C., and Gluck, M. (1995). A novelty detection approach to classification. In *Proceedings of the International Joint Conference on Aritifical Intelligence*, volume 14, pages 518–523.

Jefferys, W. H. and Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist*, 80:64–72.

Jeffreys, H. (1961). *Theory of Probability.* Clarendon Press, third edition.

Jones, S. S. and Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development*, 8:113–140.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430–454.

Kalish, C. (1998). Natural and artifactual kinds: Are children realists or relativists about categories? *Developmental Psychology*, 34(2):376–391.

Kearns, M. J. and Vazirani, U. V. (1994). *An introduction to computational learning theory*. MIT Press.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Harvard University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press.

Koenderink, J. J. and van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Lakoff, G. (1972). Hedges: A study in meaning criteria and logic of fuzzy concepts. In *Papers from the Eight Regional Meeting, Chicago Linguistic Society*, pages 183–228. Chicago Linguistic Society.

Lakoff, G. (1986). *Women, fire, and dangerous things: What categories tell us about the nature of thought*. University of Chicago Press.

Landau, B., Smith, L., and Jones, S. (1997). Object shape, object function and object name. *Journal of Memory and Language*, 36(1):1–27.

Levi, E. H. (1949). *An introduction to legal reasoning*. University of Chicago Press.

Lowe, D. G. (1985). *Perceptual organization and visual recognition*. Kluwer Academic Publishers, Boston, MA.

MacKay, D. (1992). Bayesian interpolation. *Neural Computation*, 4:415–447.

Macnamara, J. T., editor (1982). *Names for things: a study of human learning*. MIT Press.

Markman, E. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Markson, L. and Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, 385:813–815.

Marr, D. C. (1982). *Vision: A computational investigation into the human reprrepresentation and processing of visual information*. W. H. Freeman.

Medin, D. L. and Florian, J. E. (1992). Abstraction and selective coding in exemplar-based models of categorization. In Healy, A. F., Kosslyn, S. M., and Shiffrin, R. M., editors, *From learning processes to cognitive processes: Essays in honor of William K. Estes*. L. Erlbaum.

Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2):254–278.

Medin, D. L. and Ortony, A. (1989). Psychological essentialism. In Vosniadou, S. and Ortony, A., editors, *Similarity and analogical reasoning*, pages 179–195. Cambridge University Press, Cambridge.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification. *Psychological Review*, 85:207–238.

Mervis, C. B. and Crisafi, M. A. (1982). Order of acquisition of subordinate, basic, and subordinate level categories. *Child Development*, 53:258–266.

Mervis, C. B. and Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32:89–115.

Miller, K. and Gelman, R. (1983). The child's representation of number – a multidimensional-scaling analysis. *Child Development*, 54(6):1470–1479.

Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*, 21(1):55–100.

Minka, T. P. and Picard, R. W. (1997). Interactive learning using a 'society of models'. *Pattern Recognition*, 30(4).

Mitchell, T. M. (1979). *Version spaces: An approach to concept learning*. PhD thesis, Electrical Engineering Department, Stanford University.

Mitchell, T. M. (1980). The need for biases in learning generalization. Technical report, Computer Science Department, Rutgers University, New Brunswick, NJ 08904.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, 18:203–226.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Muggleton, S. Learning from positive data. Under review at *Machine Learning*.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92:289–316.

Murphy, G. L. and Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, 21:1–20.

Narayanan, S. and Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 20. L. Erlbaum.

Nisbett, R., Krantz, D., Jepson, C., and Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90:339–363.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43:25–53.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., and Glauthier, P. (1994a). Comparing models of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1964). *Memory and Cognition*, 22(3):352–369.

Nosofsky, R. M. and Palmieri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin and Review*, 5(3):345–369.

Nosofsky, R. M., Palmieri, T. J., and McKinley, S. C. (1994b). Rule-plus-exception model of classification learning. *Psychological Review*, 101:53–79.

Oaksford, M. and Chater, N., editors (1998). *Rational models of cognition*. Oxford University Press.

Osherson, D. N. (1978). Three conditions on conceptual naturalness. *Cognition*, 6:263–289.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97(2):185–200.

Osherson, D. N., Stob, M., and Weinstein, S. (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists*. MIT Press.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: An application to face detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 130–136, San Juan, PR.

Pearl, J., editor (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Petsche, T., Marcantonio, A., Darken, C., Hanson, S. J., Kuhn, G. M., and Santoso, I. (1996). A neural network autoassociator for induction motor failure prediction.

In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 924–930. MIT Press.

Pinker, S. (1984). *Language learnability and language development*. Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT Press.

Pinker, S. (1991). Rules of language. *Science*, 253:530–535.

Pinker, S. (1995). Language acquisition. In Gleitman, L. R. and Liberman, M., editors, *Language: An invitation to cognitive science, Vol. 1*, pages 135–182. MIT Press, Cambridge, MA, second edition.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497.

Popper, K. (1959). *The Logic of Scientific Discovery*. Basic Books.

Quine, W. V. (1960). *Word and object*. MIT Press.

Quine, W. V. (1969). *Ontological relativity, and other essays*. Columbia University Press.

Quine, W. V. (1995). *From stimulus to science*. Harvard University Press.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In Vosniadou, S. and Ortony, A., editors, *Similarity and analogical reasoning*, pages 21–59. Cambridge University Press, Cambridge.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7:532–547.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976a). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Rosch, E., Simpson, C., and Miller, R. S. (1976b). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2:491–502.

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.

Russell, S. J. (1986). A quantitative analysis of analogy by similarity. In *Proceedings of the International Joint Conference on Aritifical Intelligence*, volume 5.

Samuelson, L., Gasser, M., and Smith, L. (1997). Statistical regularities in input lead to a naming bias: A connectionist model of the shape bias. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 19, page 1031. L. Erlbaum.

Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*, 42:319–345.

Seung, H. S., Sompolinsky, H., and Tishby, N. (1992). Statistical mechanics of learning from examples. *Phyiscal Review A*, 45(8):6056–6091.

Shanks, D. R. and Gluck, M. A. (1994). Tests of an adaptive network model for the identification and categorization of continuous-dimension stimuli. *Connection Science*, 6(1):59–89.

Shepard, R. N. (1964). Attention and the metric structure of stimulus space. *Journal of Mathematical Psychology*, 1:54–87.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–398.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Shepard, R. N. and Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86(2):87–123.

Shepard, R. N., Hovland, C. I., and Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs*, 75(13, whole No. 517)).

Shepard, R. N., Kilpatric, D. W., and Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, 7:82–138.

Shipley, E. F. (1993). Categories, hierarchies, and induction. In Medin, D. L., editor, *The psychology of learning and motivation: advances in research and theory*, volume 30, pages 265–301. Academic Press, San Diego, CA.

Simard, P. Y., LeCun, Y., and Denker, J. (1993). Efficient pattern recognition using a new transformation distance. In Hanson, S., Cowan, J., and Giles, L., editors, *Advances in Neural Information Processing Systems*, volume 5. Morgan Kaufman.

Skyrms, B., editor (1986). *Choice and chance: An introduction to inductive logic.* Wadsworth Publishing Co., third edition.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25:231–280.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119:3–22.

Sloman, S. A. and Rips, L. J., editors (1998). *Similarity and Symbols in Human Thinking.* MIT Press.

Smith, E. E. (1995). Concepts and categorization. In Smith, E. E. and Osherson, D. N., editors, *Thinking: An invitation to cognitive science, Vol. 3*, pages 3–34. MIT Press, Cambridge, MA, second edition.

Smith, E. E., Langston, C., and Nisbett, R. E. (1996). The case for rules in reasoning. *Cognitive Science*, 16(1):1–40.

Smith, E. E. and Medin, D. L. (1981). *Categories and concepts*. Harvard University Press.

Smith, E. E., Patalano, A. L., and Jonides, J. (1998). Alternative strategies of categorization. In Sloman, S. A. and Rips, L. J., editors, *Similarity and Symbols in Human Thinking*. MIT Press.

Smith, E. E. and Sloman, S. A. (1994). Similarity- vs. rule-based categorization. *Memory and Cognition*, 22:377–386.

Sommers, F. (1963). Types and ontology. *Philosophical Review*, 72:327–363.

Spelke, E. (1995). Initial knowledge: Six suggestions. *Cognition*, 50:433–447.

Stern, J. J. (1991). *Similarity-based Likelihood Judgment*. PhD thesis, Department of Brain and Cognitive Sciences, MIT.

Sun, R. (1995). Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence*.

Sung, K. and Poggio, T. (1994). Example-based learning for view-based human face detection. A.I. Memo 1521, Artificial Intelligence Lab, MIT, Cambridge, MA 02139.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 3–9. MIT Press.

Tenenbaum, J. B. (1997a). A bayesian framework for concept learning. In *Proceedings of the Interdisciplinary Workshop on Similarity and Categorisation*.

Tenenbaum, J. B. (1997b). Making sense of typicality: What makes a good example? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 19, page 1069. L. Erlbaum.

Tenenbaum, J. B. (1998). Mapping a manifold of perceptual observations. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Todorov, E. (1998). *Studies of Goal-Directed Movements*. PhD thesis, Department of Brain and Cognitive Sciences, MIT.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

Tversky, A. and Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, 89:123–154.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76:105–110.

Tversky, A. and Kahneman, D. (1983). Extensions versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315.

Ullman, M., Corkin, S., Coppola, M., Hickok, G., H., G. J., Koroshetz, W. J., and Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, 9:289–299.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer Verlag.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12:129–140.

Wasserman, E. A. (1995). The conceptual abilities of pigeons. *American Scientist*, 83:246–255.

Watanabe, S. (1960). Information-theoretical aspects of inductive and deductive inference. *IBM Journal of Research and Development*, pages 208–231.

Watanabe, S. (1969). *Knowing and guessing: a quantitative study of inference and information*. Wiley.

Watanabe, S. (1985). *Pattern recognition: human and mechanical*. Wiley.

Waxman, S. R. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, 5:123–150.

Weiss, Y. (1998). *Bayesian motion estimation and segmentation*. PhD thesis, Department of Brain and Cognitive Sciences, MIT.

Wexler, K. and Manzini, R. (1987). Parameters and learnability in binding theory. In Roeper, T. and Williams, E., editors, *Parameters and linguistic theory*. Riedel.

Whewell, W. (1858). *Novum organon renovatum*. J. W. Parker and Son.

Witkin, A. P. and Tenenbaum, J. M. (1983). On the role of structure in vision. In Beck, J., Hope, B., and Rosenfeld, A., editors, *Human and Machine Vision*, pages 481–543. Academic Press.

Wittgenstein, L. (1953). *Philosophical investigations*. Macmillan.