THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we have read the dissertation

prepared by Colin Reimer Dawson

Inference in Infants and Adults

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy.

________________________________________ Date: 10/01/2010
LouAnn Gerken

________________________________________ Date: 10/01/2010
Rebecca Gómez

________________________________________ Date: 10/01/2010
Mary Peterson

________________________________________ Date: 10/01/2010
David Temperley

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission
of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend
that it be accepted as fulfilling the dissertation requirement.

________________________________________ Date: 10/01/2010
Dissertation Director: LouAnn Gerken
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: Colin Reimer Dawson
ACKNOWLEDGEMENTS

I would like to acknowledge the following, non-exhaustive, list of (in some cases, inexhaustible) people:

- LouAnn Gerken, my advisor, for her support and guidance from Day 1 of graduate school.
- The rest of my committee, Rebecca Gómez, Mary Peterson and Davy Temperley, for their insightful commentary and suggestions about previous drafts.
- Kathie Barnes, for teaching me the nuts and bolts of doing Bayesian modeling, and for helping me work out the details of the model described herein.
- Juliet Minton and Brianna McMillan, for entropy-deflection, and generally making the Tweety Lab go.
- Kailey Tucker, Niki Szivek, as well as Juliet and Brianna, for recruiting and running participants.
- All the parents, infants, and students who volunteered to participate in experiments.
- Beth Owens, for keeping me free of red-tape mummification.
- Peter Richtsmeier, for turning me on to \LaTeX, and pointing me to this dissertation template.
- All the creators of the above for the creation of the above.
- John Bear, for his \texttt{R} expertise.
- All my friends and colleagues in the Psychology and Statistics programs, for being idea bouncers and bouncees, but even more for helping with the unwinding.
- My dogs, Obie and Raisin (well, they think they’re people), for helping keep me grounded, and being completely indifferent to the whole dissertation thing. And, of course:
- My parents, for their unending support and encouragement, and generally setting tremendous examples to which someday I hope to live up (or is it “up to which someday I hope to live”? That’s one of those examples I don’t quite have down yet.)
- My wife, Sarah, for more than I can possibly describe.
## Table of Contents

### List of Figures ................................................. 8

### List of Tables ............................................... 9

### Abstract ...................................................... 10

### Chapter 1 Philosophical and Mathematical Foundations . 12

1.1 The Purpose of This Dissertation ......................... 12

1.1.1 What Forms Does Inference Take? ....................... 13

1.1.2 What is the Role of Experience? ....................... 14

1.2 Navigating This Dissertation ............................ 14

1.3 Deduction, Induction and Abduction ..................... 16

1.3.1 Deductive Inference .................................. 16

1.3.2 Inductive Inference .................................. 17

1.3.2.1 Probabilistic Induction .......................... 17

1.3.3 Abductive Inference .................................. 18

1.3.3.1 Probabilistic Abduction and Bayes’ Theorem ....... 20

1.3.3.2 Explaining Away .................................. 23

1.3.4 Machine Learning Models of Inference ................ 28

1.3.4.1 Discriminative Methods ......................... 29

1.3.4.2 Probabilistic Generative Methods ................. 32

1.3.4.3 Discriminating Generative and Discriminative Genera 35

1.3.5 Interim Summary ...................................... 37

1.4 Nativism and Domain-Specificity ......................... 38

1.4.1 The “Poverty of the Stimulus” Argument .............. 39

1.5 Summary ................................................... 42

### Chapter 2 A Review of Relevant Recent Empirical Research 44

2.1 Statistical Learning in Infants .......................... 44

2.2 Abductive Reasoning and Explaining Away .............. 46

2.3 Learning Abstract Rules .................................. 50

2.4 Adaptive Perception and Learning ...................... 52

2.5 Possible Adaptations in Rule Learning .................. 54

2.5.1 The Principle of Pitch Proximity, or “Smoothness”, in Music 56

2.6 Empirical Research Reported in This Dissertation .... 57
TABLE OF CONTENTS – Continued

CHAPTER 3 EXPERIMENT 1 .......................................................... 59
  3.1 Introduction ........................................................................... 59
  3.2 Methods ............................................................................. 62
    3.2.1 Participants .................................................................. 62
    3.2.2 Materials and Procedures .............................................. 62
      3.2.2.1 Procedures: Context Phase ..................................... 63
      3.2.2.2 Materials: Context Phase ....................................... 64
      3.2.2.3 Procedures: Grammar-Learning Phase ..................... 65
      3.2.2.4 Materials: Grammar-Learning Phase ....................... 66
  3.3 Results ................................................................................ 67
    3.3.1 Number Correct ........................................................... 68
      3.3.1.1 Full Sample ....................................................... 68
      3.3.1.2 Trimmed Sample ............................................... 70
    3.3.2 Discrimination Scores .................................................. 71
      3.3.2.1 Full Sample ....................................................... 71
      3.3.2.2 Trimmed Sample ............................................... 72
  3.4 Discussion ......................................................................... 72

CHAPTER 4 A BAYESIAN LEARNING MODEL FOR EXPERIMENT 1 .... 75
  4.1 Model Definition ............................................................... 75
  4.2 Simulations and Evaluation .................................................. 77
  4.3 Discussion ......................................................................... 78
  4.4 Modeling Details ............................................................... 81
    4.4.1 Model Definition ....................................................... 81
    4.4.2 Sampling Procedure .................................................. 82

CHAPTER 5 EXPERIMENT 2 .......................................................... 84
  5.1 Introduction ....................................................................... 84
  5.2 Methods ............................................................................ 87
    5.2.1 Participants .................................................................. 87
    5.2.2 Materials and Procedures .............................................. 87
      5.2.2.1 Context Phase: Materials ..................................... 87
      5.2.2.2 Context Phase: Procedures .................................... 88
      5.2.2.3 Grammar-Learning Phase: Materials ..................... 89
      5.2.2.4 Grammar-Learning Phase: Procedures ................... 90
  5.3 Results ................................................................................ 91
    5.3.1 Exclusion Criteria ........................................................ 92
    5.3.2 Planned Analyses .......................................................... 93
      5.3.2.1 4-month-olds ..................................................... 94
      5.3.2.2 7-month-olds ..................................................... 96
    5.3.3 Interim Discussion and Post Hoc Analyses .................... 97
      5.3.3.1 4-month-olds ..................................................... 98
      5.3.3.2 7-month-olds ..................................................... 99
TABLE OF CONTENTS – Continued

5.3.3.3 Comparison Across Ages ........................................ 100
5.4 Discussion ............................................................... 102

CHAPTER 6 EXPERIMENT 3 .............................................. 106
  6.1 Introduction .......................................................... 106
  6.2 Methods ............................................................... 108
    6.2.1 Participants ..................................................... 108
    6.2.2 Materials and Procedures .................................... 109
      6.2.2.1 Procedures and Materials: Context Phase ............. 109
      6.2.2.2 Procedures: Grammar-Learning Phase ............... 109
      6.2.2.3 Materials: Grammar-Learning Phase ................. 109
  6.3 Results ............................................................... 111
    6.3.1 Number Correct .............................................. 112
      6.3.1.1 Full Sample ............................................ 113
      6.3.1.2 Trimmed Sample ....................................... 113
    6.3.2 Discrimination Scores ...................................... 114
      6.3.2.1 Full Sample ............................................ 114
      6.3.2.2 Trimmed Sample ....................................... 115
    6.3.3 Comparison to Experiment 1 ................................ 115
      6.3.3.1 Number Correct ...................................... 116
      6.3.3.2 Discrimination Scores ................................ 117
  6.4 Discussion ............................................................ 119

CHAPTER 7 GENERAL DISCUSSION AND FUTURE DIRECTIONS .... 123
  7.1 Summary and Conclusions ........................................ 123
  7.2 Questions, Limitations and Ambiguities ......................... 124
    7.2.1 The Repetition Distribution ................................ 124
    7.2.2 By What Path Does Smoothness Influence Rule-Learning? . 127
    7.2.3 Conditional vs. Marginal Interval Distributions ........... 129
    7.2.4 Blocking and Generalization Over Vocabularies .......... 129
  7.3 Broader Open Questions ......................................... 130
    7.3.1 The Role of Surprise ...................................... 130
    7.3.2 The Connectionism Connection ................................ 132

REFERENCES ................................................................. 134
## LIST OF FIGURES

| Figure 3.1 | Image of Aliens Displayed During Context Phase  | 62 |
| Figure 3.2 | Interval Distributions in Context Phase        | 63 |
| Figure 3.3 | Image of Aliens Displayed During Training and Test Phases | 65 |
| Figure 3.4 | A Schematic Representation of Melodies Used in Grammar Learning | 66 |
| Figure 3.5 | Mean Percent Correct by Block and Context      | 69 |
| Figure 3.6 | Mean Discrimination Score by Block and Context | 71 |
| Figure 4.1 | Model Likelihood Function                      | 76 |
| Figure 4.2 | Performance of the Model by Block, Condition and Parameter Settings | 79 |
| Figure 5.1 | Novelty Preferences by Context and Block       | 92 |
| Figure 5.2 | Overall Looking Time by Context and Block      | 93 |
| Figure 5.3 | Preferences by Age and Context                 | 100 |
| Figure 6.1 | A Schematic Representation of Melodies Used in Grammar Learning | 110 |
| Figure 6.2 | Mean Percent Correct by Context and Block      | 112 |
| Figure 6.3 | Mean Discrimination Score by Context and Block | 114 |
| Figure 6.4 | Mean Overall Percent Correct by Experiment and Context | 117 |
| Figure 6.5 | Mean Overall Discrimination Scores by Experiment and Context | 118 |
LIST OF TABLES

Table 1.1  Conditional Probabilities for wet lawns given rain and sprinkler state . . . . . . . 25
Table 1.2  Prior Probabilities for rain and sprinkler state . . . . . . . . . . . . . . . . . . . 25
Table 4.1  Descriptions of the Model Parameters . . . . . . . . . . . . . . . . . . . . . . . . . . 77
ABSTRACT

The human desire to explain the world is the driving force behind our species’ rich history of scientific and technological advancement. The ability of successive generations to build cumulatively on the scientific progress made by their ancestors rests on the ability of individual minds to rapidly assimilate the explanatory models developed by those who came before. But is this explanatory, model-based way of thinking limited to deliberate, conscious cognition, with the larger, unconscious portion of the workings of the mind dependent on simpler mechanisms of association and prediction, or is explanation a more fundamental drive? In this dissertation I explore theoretical, empirical and computational attempts to shed some light on this question. I first present a number of theoretical advantages that model-based learning has over its associative counterparts. I focus particularly on the inferential phenomenon of explaining away, which is difficult to account for in a model-free system of learning. Next I review some recent empirical literature which helps to establish just what mechanisms of learning are available to human infants and adults, including a number of findings that suggest that there is more to learning than mere prediction. Among these are a number of experiments suggesting that explaining away occurs in a variety of cognitive domains. Having set the stage, I report a new set of experiments, one with infants and two with adults, along with a related computational model, which provide further evidence for unconscious explaining away, and hence for some for of model-based inference, in the domain of abstract, relational pattern-learning. In particular, I find that when learners are presented with a novel environment of tone sequences, the structure of their initial experience with that environment, and implicitly the model of the environment which best accounts for that experience, influences what kinds of abstract structure can easily be learned later. If indeed learners are able to construct explanatory models of particular
domains of experience which are then used to learn the details of each domain, it may undermine claims by some philosophers and cognitive scientists that asymmetries in learning across domains constitutes evidence for an innately modular organization of the mind.
CHAPTER 1

PHILOSOPHICAL AND MATHEMATICAL FOUNDATIONS

1.1 The Purpose of This Dissertation

My goal in writing this dissertation is to shed some light on the way human infants and adults make inferences about, and learn from, their world. In particular, how do learners take advantage of their experiences to help them understand the relationships among countless environmental states that are relevant to their behavior, particularly when many of these cannot be observed directly? How can a newborn infant come to connect the speech signal available to their senses — on the surface an undifferentiated stream of arbitrary sound and visible movement — to a meaningful expression of the state of another person’s mind and/or of the shared external world? Or, even more simply, how does she come to infer what patterns of reflected light indicate the presence of her mother? Each of these cognitive abilities is a form of inference, and the system in the mind/brain that allows these inferences has to be set up through a combination of experience, on the time scale of an individual organism’s life, and developmental programming, occurring over evolutionary time. Learning occurs whenever the inference system (at whatever level of abstraction) is altered by experiences. We can even describe learning itself as a form of “higher-order” inference whose conclusions concern what other inferences should occur under what conditions in the future; and one can then ask how this higher-order inference system gets set up via experience and evolution. It is this “learning to learn” on which I will emphasize in the empirical work presented in later chapters.
1.1.1 What Forms Does Inference Take?

There are at least two broad questions to ask about the nature of human learning and inference. The first is a question of what kinds of inference actually take place – what are the mechanisms involved? How does the cognitive system translate transient sensory input into new representations, and how does this translation process change with repeated input? Although a great deal of interesting inference takes place at an unconscious level, and hence is not what would normally be described as a form of “reasoning”, useful insights about both deliberate and unconscious inference can be gained by examining the formal systems of inference that have been developed to characterize explicit reasoning. Although the components of these formal systems (preconditions, implications and consequences) are traditionally logical, propositional statements, at a fundamental level, inference occurs whenever one piece of information is derived from another, of whatever form. The input to the reasoning process, what logicians would call premises, can potentially arise from preexisting knowledge, as well as present experience. To begin with, I will review the three forms of logical reasoning that have been delineated in philosophy: deduction, induction and abduction. For a given learning problem, such as that of vision, language acquisition, or explicit belief formation, we as scientists will want to ask which of these forms characterize the way the human cognitive system creates new representations. This is a question of what learning mechanisms are available to the cognitive system, and how they are applied.

In studying learning in natural cognitive systems, important insights can be gained by examining attempts to construct artificial cognitive systems from scratch. A major benefit of studying these artificial systems is that in trying to build a real, functioning learning device, many of the challenges inherent in learning are forced to become explicit. Whereas abstract theories of learning may implicitly introduce intelligent homunculi to the system, leading to an infinite regress, being forced to build a machine that works prevents one from falling into such a trap. Deductive, inductive and abductive inference each have counterparts in artificial intelligence and its child discipline of
machine learning. I discuss each of these in turn.

1.1.2 What is the Role of Experience?

The second broad question facing scientists studying learning concerns why learning mechanisms are applied the way they are. Or, perhaps more accurately, whence come the systems of inference observed in humans? This is at some level a question of nativism and empiricism. In what areas and to what extent are the “premises” that feed into learning, as well as the learning mechanisms themselves, “hard-wired”? How much of the structure contained in the inference system got there as a result of experience, as opposed to biological evolution? Moreover, are there qualitatively different inference engines devoted to different kinds of information, or are there some general principles that can describe inference and learning across domains? In the parlance of cognitive science, to what extent is learning and inference domain-specific? The interaction between the nativism dimension and the domain-specificity dimension is of interest as well: Where domain-specific mechanisms do exist, how did they get that way?

1.2 Navigating This Dissertation

In this first chapter, I review the broad philosophical questions in play, and describe several theoretical frameworks used to describe various forms of inference. I also describe some of the ways in which each of these forms of inference have been employed in Artificial Intelligence applications. I will argue that one particular form of inference, namely probabilistic abductive or explanatory inference, has a number of theoretical advantages when it comes to learning about the uncertain, yet systematic, natural world. Among these advantages is the unique ability of abduction to exhibit a phenomenon known as explaining away, which I argue is an important part of inference that intuitively ought to occur. I also briefly frame the related debates over nativism vs. empiricism and domain-specificity vs. domain-generality, and discuss how they relate to the forms of learning discussed in the first part of the chapter. I suggest that one of the most prominent arguments for
nativism and domain-specificity in language, namely the Poverty of the Stimulus argument, loses much of its force when considered in light of probabilistic abduction.

I then turn in Chapter 2 to a discussion of what the evidence so far says about how, in fact, people learn and make inferences. To what extent has empirical science supported the notion that people make use of probabilistic abduction, and how pervasive is explaining away in human inference? I will also review the literature on one particular aspect of human learning, namely abstract “rule-learning” over sequences of events or objects, where a “rule” is defined by the presence of a set of relations between elements, rather than by any particular combination of elements. Because of the importance of this form of learning in the acquisition of linguistic grammar, it has been of particular interest in discussions of what learning mechanisms are needed, as well as discussions of how domain-specific these mechanisms are. I argue that, while abstract rule-learning does appear to have a domain-sensitive component, this sensitivity can plausibly be explained by the action of a model-based abductive learning mechanism, in which evidence for a potential rule is evaluated with respect to the distributions of the input that would be expected in the absence of that rule.

In the four chapters that follow, I present a set of novel empirical findings related to explaining away in rule-learning. In Chapters 3, 5 and 6, I present a set of behavioral experiments, two with adult learners and one with infant learners. I will argue that the learners in these experiments appear first to acquire an explanatory model of sequences generated in an artificial environment, and subsequently employ that model in their inferences about the presence of an abstract rule. Depending on the combination of model type and rule type, the evidence for the rule is stronger or weaker, resulting in more or less robust learning of the rule. Critically, the ease of rule-learning is affected by prior context even as the particular training examples used to demonstrate the rule are identical across contexts. In Chapter 4, I present a formal computational model of the sort of rule-learning exhibited in one of the experiments, which exhibits a relative pattern of performance which is very similar to that of the human adult learners in that experiment.

Finally, in Chapter 7, I present a summary of the empirical findings, discuss some limitations
of the experiments, and raise some broader issues to be addressed going forward.

1.3 Deduction, Induction and Abduction

Aristotle, in his *Prior Analytics*, described three distinct forms of inference relating preconditions, implications, and consequences. In cognitive terms, we can think of implications as a statement that when some part of the environment is in a particular configuration (i.e., some preconditions are met), some other environmental configuration (the consequence) will obtain (or is likely to obtain) as well. Deriving consequences from preconditions and implications involves *deduction*. When the precondition and the consequence are known, *induction* is used to derive the implication. Finally, if it is the precondition that is unknown, it can be derived from an implication and a consequence using *abduction*. Each of these forms of inference has been studied extensively in formal logic, but each can be mapped onto various psychological models of inference as well. Some philosophical background, along with psychological analogs, is discussed for each form of reasoning in turn in the remainder of this section.

1.3.1 Deductive Inference

In *deductive inference*, specific conclusions are logically entailed by general principles. Deductive arguments often have the form of the popular syllogism: “Given that all men are mortal, and given that Socrates is a man, it must be the case that Socrates is mortal”. Here, the conclusion is licensed by *modus ponens*: given the general principle that “all entities $x$ with the property $P$ have the property $Q$” (henceforth written $P(x) \Rightarrow Q(x)$), if one discovers that $P(x)$ is true, then $Q(x)$ must also be true. An equally valid form of deductive argument is *modus tollens*: given the general principle $P(x) \Rightarrow Q(x)$, if one assumes $Q(x)$ to be false, then it must be that $P(x)$ is also false. In each case, the rule “$P(x) \Rightarrow Q(x)$” must be given as true in order to license any conclusion.

It is this form of inference that is assumed to operate in the *Principles and Parameters* theory of language acquisition (Chomsky & Lasnik, 1993). On this account, the child is innately endowed with
knowledge of Universal Grammar, which specifies, for example, the principle that languages have a
tree structure over syntactic constituents, as well as a set of (originally binary) parameters which
delineate the finite set of ways in which trees can be constructed. Upon encountering grammatical
sentences that are inconsistent with a particular parameter setting, the child’s language acquisition
device deductively sets that parameter to the only remaining setting which is consistent with the
data. This sort of disconfirmation of hypotheses is fundamental to deductive reasoning: Under the
logical certainty required by deduction, no quantity of evidence consistent with a principle can ever
confirm that it holds, for it is always possible that some counterexample exists that has not yet
been encountered. On the other hand, a single counterexample can disconfirm the general principle,
when that principle is absolute: If there is a single case in which $P(x)$ is true but $Q(x)$ is false, the
principle that $P(x) \Rightarrow Q(x)$ must be false.

1.3.2 Inductive Inference

Deduction can never confirm the truth of a general principle, no matter how many consistent pieces
of evidence are observed. Proceeding from particulars to generalities is instead the domain of
inductive inference. The argument, “All snow I have seen is white; therefore all snow is white” is
inductive: the conclusion is more general — logically a stronger statement — than the premise.
A parallel linguistic example might be, “All sentences I have heard contain a verb; therefore all
sentences must contain a verb.” Each of these examples is of the form, “$P(x)$ has never been true
for any observed $x$ where $Q(x)$ is false; therefore, $P(x) \Rightarrow Q(x)$”.

1.3.2.1 Probabilistic Induction

One way to formulate induction is as a means to increase or decrease confidence that a particular
conclusion (of the form $P(x) \Rightarrow Q(x)$) is true. Intuitively, any time an instance $x$ is observed where
$P(x)$ and $Q(x)$ are both true, this constitutes evidence in favor of the principle $P(x) \Rightarrow Q(x)$. For
example, if one observes a bird which has both the property of being a swan and the property of
being white, it would make sense to increase one's belief in the generalization “all swans are white”. Though one can never conclusively prove that all swans are white, no matter how many times in a row one observes a white swan without seeing any other color (short of examining every swan in existence), it seems reasonable that observing 100 white swans should provide greater evidence for the generalization than observing 10 white swans. A natural way to formalize confidence in the truth of a proposition is to assign it a probability, which ranges from 0 (the proposition is certainly false) to 1 (the proposition is certainly true). As evidence for a generalization increases, one's confidence in its validity increases as well.

By using probabilities, the inductive problem can be reformulated as a practical problem of prediction: generalizations are useful to the extent that they enable one to make educated guesses about aspects of the environment that have not (yet) been observed. A formal cognitive system can implement inductive generalization-as-prediction by incorporating a set of functions between various features of the world, such that when one set of features is observed, other features can be predicted. For example, upon observing an object that is red and round, one might assign some probability to the object being a ball, and some other probability to the object being an apple. What probabilities to assign can be learned through experience, based on correlations between the features. A connectionist network works this way: certain nodes are active for red objects, others for round objects, and others for apples and balls. These nodes are connected with a set of variable weights which increase when the features are active together, and decrease when one is active without the other\(^1\). In formal logical terms, the network creates a probabilistic version of the rule \(P(x) \Rightarrow Q(x)\) by observing a correlation between \(P\) and \(Q\) across various \(x\)s.

### 1.3.3 Abductive Inference

Deduction uses a precondition \((P(x))\) and a rule \((P(x) \Rightarrow Q(x))\) to derive a consequence \((Q(x))\), while induction uses a precondition and a consequence to derive a rule. The remaining mode of

\(^1\)This is a vast oversimplification of connectionist networks; however, it is a general property that holds for most architectures associated with that umbrella term.
logical inference derives a precondition from a consequence and a rule. The philosopher Charles Sanders Peirce (Peirce, 1935) describes abduction in the following way:

The form of the inference . . . is this:

1. The surprising fact, C, is observed.
2. But if A were true, C would be a matter of course.
3. Hence, there is reason to suspect that A is true.

Like induction and unlike deduction, the abductive conclusion is not strictly entailed by the premises: \( P(x) \) may well be false, and \( Q(x) \) may be true for some unrelated reason; however, as with probabilistic induction, it is reasonable to increase one’s confidence in the truth of \( P(x) \) after observing \( Q(x) \). Consider the following example, courtesy of Pearl (1988). Suppose you wake one morning from a deep and restful slumber in your home in the desert, and you walk outside to discover that the grass in your front yard is wet (leaving aside for the moment why you have grass in the desert to begin with). Being a worldly sort, you have learned over the years that this wet grass phenomenon has a tendency to arise after it rains. Since you were so blissfully unconscious during the night, you don’t have any other information about whether it rained. Moreover, you live in the desert, so if you hadn’t observed the grass, you wouldn’t have suspected rain. However, if it did in fact rain, then the wet grass would be a “matter of course”, and so it seems quite reasonable to suppose that this is what happened.

A similar example is the problem of medical diagnosis. The medical diagnostician is faced with the problem of observing some set of symptoms, and making a determination about the most likely cause. The doctor comes into the situation equipped with expertise about what diseases cause what symptoms; for example, contracting the flu tends to cause a headache, a fever, and neck stiffness. Therefore, if a patient has these symptoms, this is evidence for the hypothesis that she has a case of the flu. This is not a deductive conclusion: the fact that someone is suffering from these three symptoms simultaneously does not logically entail that she has the flu (she might, for example,
have meningitis, or it may be that the headache and neck stiffness are a result of tension, and the fever has a different cause). Rather, making the diagnosis is an instance of abduction, as one reasons backward from effect to cause, using knowledge (or at least beliefs) about what kinds of observations are generated under what conditions.

1.3.3.1 Probabilistic Abduction and Bayes’ Theorem

Since abduction, like induction, cannot produce conclusions with certainty, it is sensible to think about it as a probabilistic inference process as well. Just as confidence in an inductive generalization can be mapped to a probability between 0 and 1, so too can one’s confidence in a hypothesized cause for an observed phenomenon. We usually think about probabilities as applied to events, where the probability of an event \( E \) occurring (under some specified set of conditions \( C \)) corresponds to the proportion of the time that we expect \( E \) to occur in the long run, whenever the conditions in \( C \) are present. We can write this quantity as \( P(E|C) \) (read as “the probability of \( E \) given \( C \)”). If a number of different events and conditions are possible (call them \( E_1 \) through \( E_m \) and \( C_1 \) through \( C_n \)), then a complete causal model of the world needs to assign values to \( P(E_i|C_j) \) for each possible combination of \( i \), ranging from 1 to \( m \), and \( j \), ranging from 1 to \( n \). Some events may be impossible under certain conditions; in that case, the corresponding probability is set to zero.

In order to apply the notion of probability in the other direction (i.e., what is the probability of certain conditions being present given that we have observed some event?), we need to generalize the notion of probability somewhat. After all, once an event \( E_i \) has occurred, the conditions in \( C_j \) are presumably either present or not (even though we may not know which), and thus there is something odd about asking “what proportion of the time will \( C_j \) hold if I repeat this situation indefinitely”. Instead, we need the notion of subjective probability, where a value is assigned to each of a set of “possible worlds” (even though only one of those worlds may in fact ever occur).

---

\(^2\)The use of a probability to represent a degree of belief is associated with the so-called “subjectivist” view associated with Bayesian statistics, which I describe below. Peirce himself was a frequentist (Peirce, 1935), in opposition to this interpretation, and hence would not have endorsed the foregoing discussion.
If we want to know the probability of being in some type of world (say, a world in which $C_j$ holds), then we only need to add up the probabilities of all of the (mutually exclusive) “individual worlds” of that type (one in which $C_j$ holds and $E_1$ occurs, one in which $C_j$ holds and $E_2$ occurs, etc.). That is, the marginal probability of $C_j$, which we simply write $P(C_j)$, is the sum of the joint probabilities $P(C_j \text{ and } E_i)$, often written $P(C_j, E_i)$, for each value of $i$. Now, suppose we know we are in a world where $E$ has occurred (we’ll drop subscripts for conciseness), and we want to know the probability that we are also in a world where $C$ holds (i.e., what is $P(C|E)$?). What we are really asking is, of all of the probability “stuff” assigned to possible worlds where $E$ occurs, how much of it corresponds to worlds where $C$ also occurs? In other words, out of the total (marginal) probability $P(E)$, how much comes from $P(E, C)$? This is just a matter of division:\(^3\)

$$P(C|E) = \frac{P(C, E)}{P(E)} \tag{1.1}$$

Now, this is all well and good if we have direct access to all possible joint probabilities: simply add them to get marginal probabilities, and divide to get conditional probabilities. But often it is not joint probabilities which are the most natural starting point. Recall that the causal model of the world we defined at the beginning of this section specified conditional probabilities, not joint probabilities. That is, we started with some theory (say, about the effects of various diseases on the body), and can quantify how likely each in a set of possible outcomes (e.g., symptoms) is. Unfortunately, we have conditional probabilities going in the wrong direction if what we want to do is assign a probability to a cause. In order to compute a conditional probability in the opposite direction, we first need to find the joint probability. Notice that in deriving 1.1, we didn’t appeal to anything inherently asymmetric about the relationship between $E$ and $C$. We simply began with “atomic” possible worlds, each assigned a joint probability, and computed a proportion to get a conditional probability. It would be equally valid to compute the proportion in the other direction.

\(^3\)The concepts being presented in this and the next section can be understood without following the algebraic details. I provide step-by-step examples only to show the interested reader that the intuitive conclusions really do follow from the math.
By switching the places of $C$ and $E$ in equation 1.1, we get:

$$P(E|C) = \frac{P(E,C)}{P(C)} \quad (1.2)$$

It is then a simple algebraic maneuver to solve for $P(E,C)$ in terms of $P(E|C)$:

$$P(E,C) = P(E|C)P(C) \quad (1.3)$$

It’s clear now that we need to be given one more piece of information if we want to compute $P(C|E)$ — namely the marginal probability $P(C)$. Previously we computed marginal probabilities by summing over joint probabilities, but since now we are trying to derive the latter, that approach is no longer available. We can think of the quantity $P(C)$ as the likelihood of the conditions $C$ being present if we hadn’t made any observations — it is the prior probability of $C$. To revisit our medical example, we might ask “what is the probability that any given person has the flu?” Without identifying a particular individual, or saying anything about symptoms, this is another way of asking “how common is the flu?”. The decomposition of the joint probability $P(E,C)$ into a prior $P(C)$ and a conditional $P(E|C)$ should seem natural when probabilities are considered in terms of proportions. If we want to know, for example, “what proportion of people have the flu and a headache?” ($P(\text{flu, headache})$), we can first ask “what fraction of people have the flu?” (i.e. $P(\text{flu})$), and then, “of those, what fraction have a headache” ($P(\text{headache}|\text{flu})$). The overall fraction is simply the product of the two.

Now we have all we need to derive $P(C|E)$ from $P(E|C)$ and $P(C)$. Combining Equations 1.1 and 1.3, we have:

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)} \quad (1.4)$$

This relationship is known as Bayes Theorem. Notice that we need not be given $P(E)$ directly:
now that we are able to derive joint probabilities for each combination of $E$ and $C$, we can sum over all of these for a particular $E$ to get its marginal probability. Bringing back the subscripts, we have:

$$P(C_j|E_i) = \frac{P(E_i|C_j)P(C_j)}{\sum_{k=1}^{n} P(E_i|C_k)P(C_k)}$$ (1.5)

The right-hand-side expression consists entirely of two types of quantities: prior probabilities over conditions (or hypotheses), and conditional probabilities of observable events (data) given possible hypotheses. In the abductive setting, the data has already been observed, and so these conditional probabilities are varying over possible hypotheses, not over possible data. When used in this way, as a function over hypotheses, the “forward” probability is known as a likelihood function (since it tells us the “likelihood” of our data arising under each of a set of potential causes). Together, the prior and the likelihood are sufficient to compute the posterior probability of any given hypothesis.

### 1.3.3.2 Explaining Away

In order to engage in abductive inference, a cognitive agent must possess a model of the environment: that is, one must represent what kinds of observable data get generated for each possible state of the world. There is an essential notion of causality here that need not be present in inductive inference. While arriving at an appropriate causal model from data is a more complex problem than simply developing a good prediction method, there are valuable advantages in possessing such a model. Perhaps the most important of these is the ability to naturally extrapolate beyond what one has already observed: a causal model has the ability, in principle, to simulate new environments, and make predictions for new combinations of world states. A second advantage of a causal model is its ability to detect patterns of observations that are unlikely overall (combining across possible hidden world states), potentially signaling a novel environmental state, or an error of perception.
There is at least one other advantage of modeling causation, which is the focus of the empirical research in this dissertation. Namely, causal models possess the ability to very naturally engage in a particular form of abductive inference, familiar at an intuitive level to anyone who has ever made an explicit inference, known as explaining away.

Recall the wet grass scenario described before, but suppose now that your lawn is equipped with an automatic sprinkler system. Now there are two potential propositions, $R$: it rained, and $S$: the sprinkler system was active, either of which would render the wet grass unsurprising. As such, the wet state of the lawn (call it $W$) serves as abductive evidence for both $R$ and $S$: after observing $W$, the belief that each of $R$ and $S$ is true should increase.

Now, suppose after taking a moment to reflect on your wet lawn, you take a stroll down the block, and you happen to notice that your neighbor’s lawn is wet as well (call this fact $N$). This observation provides further support for $R$, but is by itself irrelevant to $S$ (assuming your sprinkler does not spill over to your neighbor’s yard). Moreover, a priori, $R$ and $S$ are unrelated: discovering that it rained should have no direct impact on the likelihood of the sprinkler having come on. Intuitively, however, the new observation seems to make the sprinkler hypothesis less likely than it was when only your own wet lawn had been observed. In some sense, $R$ and $S$ “compete” for the evidence contained in your own wet lawn, such that when the probability of $R$ increases, it “uses up” more of that evidence, leaving less to support $S$. The previous evidence for $S$ is explained away.

While the intuition behind explaining away should be fairly straightforward, it is instructive to see how this component of inference follows from the application of Bayes’ Theorem. To make things simple, let’s assume that both rain and the sprinkler result in a wet lawn with certainty, i.e., with probability 1. More formally, $P(W = 1|R = 1) = P(W = 1|S = 1) = 1$ (we’ll suppose a variable is equal to 1 when the proposition it represents is true, and 0 when it is false). Rain also dampens the neighbor’s lawn with probability 1, so $P(N = 1|R = 1) = 1$ as well. We’ll also suppose there is no other cause for $W$, that is, $P(W = 1|R = 1, S = 1) = 1$, but that $N$ occurs with probability 0.1 in the absence of rain, regardless of whether $S$ occurred or not. Finally, we assume
that $W$ and $N$ are independent given $R$ and $S$, so that $P(W, N|R, S) = P(W|R, S)P(N|R, S)$. To sum up, for the four possible sets of conditions (rain or not, combined with sprinkler or not), we have the following sets of conditional probabilities:

<table>
<thead>
<tr>
<th>$(R, S)$</th>
<th>$(R, S) = (0, 0)$</th>
<th>$(R, S) = (0, 1)$</th>
<th>$(R, S) = (1, 0)$</th>
<th>$(R, S) = (1, 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(W = 1</td>
<td>R, S)$</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$P(N = 1</td>
<td>R, S)$</td>
<td>0.1</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>$P(W = 1, N = 1</td>
<td>R, S)$</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1.1: Conditional probabilities for wet lawns given rain and sprinkler state

Now, since our example takes place in the desert, suppose the prior probability of rain, $P(R = 1)$, is relatively small, say 0.2, whereas the prior probability of the sprinkler running, $P(S = 1)$, is 0.5. Finally, we assume these events are independent, so that $P(R, S) = P(R)P(S)$. Together, we have joint prior probabilities for $R$ and $S$ as follows:

<table>
<thead>
<tr>
<th></th>
<th>$R = 0$</th>
<th>$R = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S = 0$</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>$S = 1$</td>
<td>0.4</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 1.2: Prior probabilities for rain and sprinkler state

Let’s calculate the \textit{a posteriori} probability of both $R$ and $S$ after observing $W$ (but before we know anything about $N$). Applying Bayes’ Theorem, we have:

\[
P(R = 1|W = 1) = \frac{P(W = 1|R = 1)P(R = 1)}{P(W = 1)} = \frac{P(W = 1|R = 1)P(R = 1)}{P(R = 1, W = 1) + P(R = 0, W = 1)}
\]
\[
P(W = 1|R = 1)P(R = 1) \\
\frac{P(W = 1|R = 1)P(R = 1)}{P(W = 1|R = 1)P(R = 1) + P(W = 1|R = 0, S = 1)P(R = 0, S = 1)} \\
= \frac{1(0.2)}{1(0.2) + 1(0.4)} \\
= \frac{0.2}{0.6} = 0.33
\]

Going from line 2 to line 3, we have that \( P(R = 0, W = 1) = P(R = 0, S = 1, W = 1) \), since in the event that it didn't rain, your lawn can be wet only if the sprinkler was on. We can compute conditional probabilities for joint events (such as \( R = 0, S = 1 \)) just as we do for simple events.

We can perform the same computation for the sprinkler:

\[
P(S = 1|W = 1) = \frac{P(W = 1|S = 1)P(S = 1)}{P(W = 1)} \\
= \frac{P(W = 1|S = 1)P(S = 1)}{P(W = 1|S = 1)P(S = 1) + P(W = 1|S = 0, R = 1)P(S = 0, R = 1)} \\
= \frac{1(0.5)}{1(0.5) + 1(0.1)} \\
= \frac{0.5}{0.6} = 0.83
\]

As expected, the probabilities of \( R \) and \( S \) both increased upon observing \( W \). What happens if we now observe \( N \)?
$P(R = 1|W = 1, N = 1) = \frac{P(W = 1, N = 1|R = 1)P(R = 1)}{\sum_{(r,s)} P(W = 1, N = 1|R = r, S = s)P(R = r, S = s)}$

$= \frac{P(W = 1, N = 1|R = 1, S = 0)P(R = 1, S = 0)}{\sum_{(r,s)} P(W = 1, N = 1|R = r, S = s)P(R = r, S = s)}$

$+ \frac{P(W = 1, N = 1|R = 1, S = 1)P(R = 1, S = 1)}{\sum_{(r,s)} P(W = 1, N = 1|R = r, S = s)P(R = r, S = s)}$

$= \frac{1(0.1) + 0.1(0.1)}{0(0.4) + 0.1(0.4) + 1(0.1) + 1(0.1)}$

$= 0.20 \div 0.24 = 0.83$

Observing the neighbor’s wet lawn adds quite a bit of evidence for rain, as expected. But what about the sprinkler hypothesis? Our intuition was that the increase in the probability of $S$ upon observing $W$ would be partially undone by the observation of $N$. Let’s crunch the numbers:

$P(S = 1|W = 1, N = 1) = \frac{P(W = 1, N = 1|S = 1)P(S = 1)}{\sum_{(r,s)} P(W = 1, N = 1|R = r, S = s)P(R = r, S = s)}$

$= \frac{P(W = 1, N = 1|R = 0, S = 1)P(R = 0, S = 1)}{\sum_{(r,s)} P(W = 1, N = 1|R = r, S = s)P(R = r, S = s)}$

$+ \frac{P(W = 1, N = 1|R = 1, S = 1)P(R = 1, S = 1)}{\sum_{(r,s)} P(W = 1, N = 1|R = r, S = s)P(R = r, S = s)}$

$= \frac{0.1(0.4) + 1(0.1)}{0(0.4) + 0.1(0.4) + 1(0.1) + 1(0.1)}$
Indeed, the neighbor’s wet lawn “explains away” some of the previous evidence for $S$. It is not quite to its a priori level, however, since it is possible that the neighbor’s lawn could be wet for a reason unrelated to rain, in which case $W$ still provides some evidence for $S$.

Consider a second example from medicine. The presence of a headache, fever and rash can be taken as evidence for the flu, meningitis, or lyme disease. Given its greater prior prevalence, the flu seems to be the most likely explanation for the symptoms. However, suppose the patient also has a bullseye-shaped rash on her thigh; a symptom of a tick bite which causes lyme disease. Now the probability of that cause increases, but in addition, the probabilities of the other two ailments go back down, as what had previously constituted evidence for them is largely explained away by the lyme disease (the reader is invited to work out this example with numbers as in the preceding scenario).

Explaining away arises very naturally in an abductive inference system; it is the direct consequence of computing the posterior probability of a hypothesis via a product of a prior and a “forward” conditional probability or likelihood. What would be required to capture this aspect of inference in a purely inductive system? That is, suppose instead of deriving posterior probabilities from priors and likelihoods, one were to attempt to learn them directly. In such a system, it would be necessary to learn a separate set of probabilities for all combinations of data patterns (e.g., my lawn is wet and the neighbor’s isn’t, both lawns are wet, etc.). As the number of data variables increases, the problem becomes exponentially more difficult.

1.3.4 Machine Learning Models of Inference

As discussed in section 1.1.1, the field of artificial intelligence has incorporated all three modes of inference at various times. In the earliest stages of AI, purely symbolic, Boolean logic algorithms
were developed to carry out automated deductive inference. These systems gave rise to automated theorem-proving computers, and discrete algorithms for playing highly constrained games such as chess. However, such methods are only useful in cases where nearly complete information about a problem domain is available. For problems where it was desirable to generalize beyond the available data, inductive and abductive methods were needed.

The extension of artificial intelligence methods beyond rule-based deductive inference systems led to the creation of a new field of *machine learning*, which borrowed much more from fields outside computer science, such as statistics and engineering, than did traditional AI. Today, machine learning methods can be roughly divided into *discriminative* and *generative* methods, each with a different approach to learning about the structure of the environment. Loosely speaking, discriminative methods carry out inductive inference, inferring heuristics to predict one feature of the environment from another. In other words, the system learns to *discriminate* among patterns of data based on what hidden features they are associated with. Generative methods, on the other hand, carry out abductive inference: equipped with a model of how observable data arises from a set of causes, they are able to make inferences concerning which causes are responsible for a given data point. Because they represent the causal structure of the environment, they are capable of *generating* new, hypothetical data, a capacity outside the purview of discriminative methods.

1.3.4.1 Discriminative Methods

**Neural Networks** One of the earliest artificial inductive systems capable of generalizing beyond training data was the perceptron learning algorithm (Rosenblatt, 1958), based on the McCulloch-Pitts “logic gate” model of a neuron (McCulloch & Pitts, 1943). This simple algorithm assigns binary classes to arbitrarily high-dimensional data by learning a set of linear weights governing the relationship between each input and output feature. While this method is powerful for certain classes of classification problem, the perceptron cannot learn any distinction for which the two classes cannot be separated by a linear boundary — i.e., an $n - 1$-dimensional hyperplane in an
$n$-dimensional feature space (a zero-dimensional dividing point when space is a 1D line, a line in 2D space, etc.). Even a simple problem in two binary-valued dimensions such as the Exclusive Or (XOR) logic gate could not be learned by the perceptron algorithm.

After Minsky and Papert (1969) pointed out this important limitation of the single-layer neural network, the entire neurally-inspired architecture fell out of favor for nearly two decades before Rumelhart and McClelland (1986) developed a workable method for training networks with multiple layers of McCulloch-Pitts-style neurons, augmented with continuous, rather than binary threshold, activation functions. By including an intermediate level of representation, composite features, in which space the classification problem was linearly separable, could be learned. For example, the XOR relation over inputs $x_1$ and $x_2$ can be decomposed into $(x_1 \ OR \ x_2) AND (NOT(x_1 \ AND \ x_2))$. Each of these other relations, AND, OR and NOT, are linearly separable problems, and hence a network which represents composite features such as $(x_1 \ OR \ x_2)$ in an intermediate layer can successfully learn the more complex relationship.

Kernel Classifiers  The way that the multi-layer perceptron solves the linear separability problem is by projecting data points from one feature space to another, potentially higher-dimensional, space, in which the classes to be discriminated are linearly separable. Kernel-based methods, such as the Support Vector Machine (SVM) Cortes and Vapnik (1995); Vapnik (1995), also project data into a new feature space in order to achieve separability, however their approach is somewhat different: Rather than applying a predefined set of basis functions which are transformations of the original features, kernel methods create features which are a function of individual training points. Whereas traditional neural networks classify new data points based on how well they resemble the “central essence” of each category, kernel methods compare new data to a subset of individual training cases directly. In this way, the latter bear a resemblance to exemplar-based categorization systems developed in psychology (Medin & Schaffer, 1978; Nosofsky, 1984, 1986). A major advantage of this approach is that useful learning can take place when the set of features being represented is
greater than the number of available observations for training. SVMs and their descendants today routinely achieve state-of-the-art performance in benchmark classification tasks.

Probabilistic Extensions Although in their simplest forms the above discriminative algorithms are deterministic, probabilistic analogs to each have been developed as well. Probabilistic discriminative models extend their deterministic counterparts by modeling the posterior distribution of a class given the features of a data point. In the case of the multilayer perceptron, as trained using the backpropagation algorithm (Rumelhart & McClelland, 1986), the activations of output nodes corresponding to classes have been shown to approximate the posterior probability of class membership (Ruck, Rogers, Kabrisky, Oxley, & Suter, 1990; Hampshire & Perlmutter, 1990), provided the distribution of categories in the training set mirrors the true distribution in the world. Support vector machines as originally formulated are not interpretable in terms of posterior probabilities, but variations such as the relevance vector machine (Tipping, 2001) do produce this probabilistic output. Unlike generative models, however, discriminative methods attempt to learn posterior distributions directly, rather than via the joint distribution of classes and features.

Unsupervised Methods In some cases, one wishes to make predictions about variables which are unobserved even in “training” data. Methods such as perceptrons (single and multi-layer) and support vector machines are supervised methods, as they require the input of a “teacher” who gives the correct labels for a subset of data. However, there are standard discriminative methods for unsupervised learning as well, where a set of classes can be learned from unlabeled data. The idea is that observations that belong to the same category should be more similar to each other than they are to members of another category. Provided a sensible measure of similarity is defined, “clusters” of observations can be inferred from data, and new data points assigned (either probabilistically or in a “winner-take-all” manner) to a cluster. Algorithms such as k-means clustering (MacQueen, 1967) accomplish this by choosing cluster centers and cluster assignments such that the aggregate
distance from observations to their respective centers is minimized. This kind of clustering can be implemented in a neural network whose output nodes represent cluster membership competitively, i.e., where activation in one output node inhibits activation in the others. In traditional \( k \)-means clustering, the value of \( k \) (i.e., the desired number of clusters) must be specified in advance; often an undesirable requirement when there is little theoretical guidance for choosing \( k \). In many engineering applications, trial and error is used to determine a good value of \( k \). A slightly more principled approach involves running the algorithm several times on the same data, and assigning a score to each run in which higher values of \( k \) are penalized (since by absolute measures, one can always improve the fit to the data by including more clusters). One can then choose the value of \( k \) which strikes the best balance between model fit and simplicity.

1.3.4.2 Probabilistic Generative Methods

The defining characteristic of generative learning models is that they attempt to represent the entire joint distribution over all variables of interest, assigning a probability to every combination of values. As described in section 1.3.3.1, the posterior distribution over possible values of an unobserved variable (such as a category label), given the observed data, can then be calculated from the full joint distribution by computing a marginal probability of the observed data and dividing the joint by the marginal.

Although in principle learning the full joint distribution directly would provide the most accurate model of the environment for a given set of variables (the choice of what variables to model in the first place is a separate, and difficult, question), the amount of data required to reliably do so increases dramatically as the number of variables increases. This need for massive data sets in order to adequately cover the data space is often referred to as the curse of dimensionality (Bellman, 1957, 1961) and is a major motivation for using discriminative methods, such as Support Vector Machines, that allow useful work to be done with relatively few data points. By imposing a causal structure to the set of variables, however, one can restrict which variables can be directly dependent
on each other, and the curse of dimensionality can be mitigated.

**Graphical Models** One important way in which imposing structure on the variables simplifies matters is through the introduction of *conditional independence* relationships. By specifying that variables $A$ and $C$ are conditionally independent given the value of $B$, we stipulate that, when the value of $B$ is known, learning more about $A$ will not tell us anything about $C$, and vice versa. This reduces the problem of estimating the joint distribution of $A$, $B$, and $C$ to the problem of estimating the joint distribution of $A$ and $B$, along with the distribution of $B$ and $C$. This may not sound easier, but imagine that each of $A$, $B$ and $C$ can take on 10 values. The three-way joint distribution involves 1000 combinations, whereas each two-way distribution involves only 100 combinations, and hence the magnitude of the problem is reduced by a factor of five. The savings will be even more dramatic in problems with greater numbers of variables and greater numbers of possible values.

The conditional independence structure among a set of variables can be represented using a *graphical model*, where each variable is represented as a node, and each direct dependence is represented as an edge connecting two nodes. One particular variety of graphical model, known as a *Bayes Net* (see, e.g. Pearl, 1988), is especially useful in formalizing abductive inference, as the conditional independence structures it contains lend themselves to a causal interpretation. In a Bayes Net, pairs of variables are connected via directional edges. In causal terms, a connection from $A$ to $B$ can be thought of as a statement that $A$ has an influence on $B$. If $B$ also leads to $C$ but $A$ does not directly lead to $C$, then $A$ and $C$ are conditionally independent given $C$ (since $A$ only influences $C$ indirectly, through $B$). The full joint distribution over all variables in the graph can be written as a product of prior marginal distributions over the variables with no edges leading to them, and the conditional distributions of the remaining variables given the set of variables leading directly to them.

An interesting property of Bayes Nets concerns the role of directionality in determining con-
ditional independence structure. In most cases, two nodes are conditionally independent given a third if (1) they lead to each other through the third variable, or (2) they are common effects. However, the case of two common causes is distinct: Two variables that both lead to a third (but are otherwise not connected) are independent when the third variable is not observed. As soon as the third variable is observed, the independence is broken, and learning about one cause affects inferences about the other. This is precisely what happens in the explaining away situation: as long as the common effect, $C$, of two variables, $A$ and $B$, is unknown, learning about $A$ may well be uninformative about $B$ (as with the sprinkler and rain in section 1.3.3.2). However, as soon as the state of $C$ is known (providing information about $A$ and $B$), it “unblocks” the path connecting $A$ and $B$, and learning about one is informative about the other (as when learning about the rain affected inferences about the sprinkler in light of the wet lawn).

The Naïve Bayes Classifier

One of the simplest forms of Bayes Net is the Naïve Bayes Classifier, which, as the “classifier” part of its name suggests, is often used to assign observations to categories. However, as a generative model, it can also be used to synthesize a new data set. The intuition behind this method is that the observable features of an object are conditionally independent of each other, given the category label.

Consider the task of classifying people by gender based on their height and their hair length. In the absence of information about gender, height is a negative predictor of hair length; however, we might reasonably assume that tall females are no more or less likely to have long hair than short females, and similarly for males. The corresponding graphical model would consist of a gender node, which leads to both height and hair length nodes, but where height and hair length are not directly connected. The joint probability of any given person being a male who is over six feet tall with hair less than six inches long is then simply the prior probability of being male, times the conditional probability of a male being six feet tall, times the conditional probability of a male having hair less than six inches long. The same joint probability can be calculated for a female
with the same attributes. The marginal probability of having those attributes is then the sum of these two joint probabilities. Finally, the posterior probability of being male given the observed attributes is the corresponding joint probability divided by the marginal probability. With the conditional independence assumption in place, one only needs enough training data to estimate the univariate distributions of gender, height given gender, and hair length given gender, in order to produce a working classifier. In contrast, without imposing this structure on the data, one would need enough data to cover a trivariate distribution.

**Hidden Markov Models** Another version of a Bayes Net is the Hidden Markov Model (HMM), used for data with a temporal or other sequential component. Here, a hidden process is assumed to evolve over time such that the state at time $t$ depends only on the state at time $t - 1$. Moreover, for an observable sequence produced by the hidden process, the state at time $t$ is assumed to depend on the state of the hidden process only at the concurrent time step. Thus, given a set of transitional probabilities of evolving from one hidden state to another, along with a set of emission probabilities of each observable state obtaining when the process is in a particular hidden state, it is possible to estimate the probability of any given hidden sequence, given an observed sequence. There are standard algorithms that can efficiently learn transitional and emission probabilities from labeled or unlabeled training data (the latter is an unsupervised learning problem, the standard solution to which makes use of a version of the Expectation-Maximization algorithm for probabilistic clustering).

**1.3.4.3 Discriminating Generative and Discriminative Genera**

The defining difference between discriminative and generative models is the ability of the latter to simulate new data. In some cases this may be an end in itself. However, even if the generation of a realistic distribution of novel data is not a goal, the ability to do so confers a number of other advantages, some of which were alluded to at the start of section 1.3.3.2.
By developing a “theory” about the processes that created the data, one can more naturally generalize beyond what has already been observed. Although all machine learning models are designed with some generalization in mind, for many discriminative models this generalization is limited to interpolation. In contrast, a learner which is guided by a theory, especially one which is hierarchical in nature, can make useful predictions in an entirely novel context, by drawing on information at a higher level of abstraction, which may connect the new context to one about which specific details are known.

Another practical advantage of a generative model is its ability to detect novel situations. By representing joint distributions over variables of interest, it is possible to compute the marginal likelihood of a particular set of variables taking on some observed values, even without making a commitment to particular values of the unobserved variables. In cases where this marginal likelihood is low, it may be desirable to add a new context to one’s theory. A discriminative model, even a probabilistic one, on the other hand, has no notion of marginal likelihood, and simply produces a posterior distribution over the unobserved variables in which the likelihood of the data is already normalized out.

For present purposes, the most important advantage of the generative approach is the natural emergence of explaining away behavior. In the formalism of a Bayes Net, whenever an observed node has more than one path leading to it (i.e., it has multiple “influences”) learning about the state of one of its influences affects inferences about the other. If the graph is correctly specified, this is quite a rational and desirable consequence; however it is quite difficult, and at the very least unnatural, for discriminative models to produce it. Consider the task faced by the feed-forward neural network charged with learning that the neighbor’s grass being wet is evidence against the sprinkler having been on, but only when one’s own grass is wet: it would need a compound feature representing the joint state of both lawns. While this may not be a big deal in this simple example, as the number of potential causes and effects increases, the number of compound features necessary to produce all desired inferences balloons dramatically. In contrast, a system with a suitably structured network
of variables produces such induced relationships quite effortlessly.

1.3.5 Interim Summary

In this section I have presented three forms of inference — deduction, induction and abduction — that have their roots in classical logic. While each was originally stated for absolute, propositional content, all three can be adapted to describe cognitive inference in a variety of domains, from perception, to language-learning, to explicit reasoning. Indeed, each of these inference modes has been used in the field of artificial intelligence, the high-level goal of which is generally to develop explicit computational methods to reproduce the kinds of inferences that humans produce quickly and easily. While deductive and inductive methods have older and more firmly established cognitive theories describing how they might occur in humans, I have argued, as others have in recent years, that it may be necessary, or at least more natural, to model many aspects of human cognition with abductive formalisms, particularly in cases where multiple causes can account for the same observed data. In these situations, it is desirable to allow these hypotheses to compete for the evidence, even in cases where they may be independent of each other a priori. In an abductive system, this competition occurs via explaining away, a phenomenon which is difficult to account for with other forms of reasoning, without building in ad hoc heuristics. The framework of Bayesian probabilistic inference, particularly as realized in graphical models, provides a principled method for capturing abductive inference, and produces explaining away behavior without stipulating it explicitly, maintaining a priori independence of multiple causes when such independence is warranted.

As discussed in the opening to this chapter, the second major component of this dissertation concerns the why (or whence) question of learning mechanisms. How much of the machinery which supports inference must be provided in advance, independent of experience, by biology? How much of that machinery can itself be constructed by learning? To what extent, and in what ways, does the innate endowment differentiate a priori among the various domains in which inference takes place? Clearly, powerful constraints on learning must be present for any cognitive system that hopes to
navigate the infinite space of inferences that can potentially be made from finite experience; how specific and invariant must these constraints be? I turn now to these questions of nativism and the domain-specificity of learning.

1.4 Nativism and Domain-Specificity

Few if any debates in cognitive science (or its parent branch of philosophy) are older, more passionately argued over, or more culturally significant, than that between nativism and empiricism — the so-called “nature vs. nurture” debate. In the popular imagination, the dichotomy is cast as one between a view that everything in the mind follows from the genes, and the view that the mind is a fully general “blank slate”, all of its content constructed from sensory experience. Of course, as with most debates popularly cast as dichotomies, both extremes are straw men. The information contained in the genome is meaningless without interaction with the environment, even for the construction of our “hardware”, let alone the contents of the mind. And even the most committed behaviorist must acknowledge that the ability to take advantage of sensory experience is firmly rooted in one's biological makeup: a puppy and a human baby, raised with identical environmental input, would not grow up with the same chalk-marks on their slates. As Chomsky (1980) put it, “the question is not whether innate structure is a prerequisite for learning, but rather what it is.”

It makes little sense to ask, about some content or feature of the mind, “is it due to genes or to learning?” A more reasonable question concerns the degree of invariance of the products of experience. Although a nativist cannot deny that environmental input at some level is needed to produce every feature of the mind, it would be \textit{a priori} defensible to claim that the resulting feature would be substantially the same for a wide range of reasonable (in some sense, “non-pathological”) inputs. A distinct but related question is to what extent general-purpose learning mechanisms can be invoked across a wide variety of domains, and to what extent the biological endowment provides for a mind which invokes different machinery for different kinds of inputs. In short, how
domain-specific is learning, and in what ways? These two axes\(^4\), invariance and domain-specificity, are commonly conflated. However, they are, in principle, separable. To the extent that experiences are internalized in domain-specific ways, we can ask to what extent these differences are themselves experience-invariant. Domain-specificities could, after all, be produced through experience. This possibility has been explored both philosophically (Karmiloff-Smith, 1992) and computationally (Jacobs, 1997, 1999) in recent decades.

1.4.1 The “Poverty of the Stimulus” Argument

Somewhat fittingly, cognitive scientists who aim to defend positions traditionally associated with nativism — namely those in the “invariant” and “domain-specific” corner of philosophical space — often prefer to rely on primarily logical, rather than primarily empirical, arguments. This spirit is embodied in the Poverty of the Stimulus (PotS) argument for an innate language faculty, which constitutes an argument for a highly invariant and domain-specific language acquisition apparatus, taking the form of a proof whose premises are meant to be intuitively reasonable, rather than data-based. The argument, originally due to Chomsky, is summed up nicely by Laurence and Margolis (2001), whom I paraphrase below:

\(^4\)These are not, of course, one-dimensional continua; but I adopt the present terminology in the interest of visualization.
The Poverty of the Stimulus Argument:

1. The correct structure of language is not uniquely determined by the child’s input.

2. The correct structure is also no simpler or more “natural”, in any general way, than the alternatives.

3. The kinds of input that would uniquely specify the correct structure are not generally available to children.

4. By 1, 2 and 3, a learner without language-specific knowledge or biases could not infer the correct grammar.

5. Children in fact acquire the correct grammar for the language(s) they are exposed to.

6. By 5 and 6, children must have language-specific knowledge and biases.

Although the PotS argument was formulated for language acquisition, its insights apply to other inferential problems as well. It is a general fact that any finite data set (e.g., the set of sentences encountered by a child) is compatible with multiple generalizations about just which new data points (e.g., possible novel linguistic utterances) belong in the same category (e.g. grammatical sentences), provided categories are allowed to comprise arbitrary subsets of the set of possible data points. When the latter set is infinite, as it is for linguistic utterances, it contains an infinite number of distinct subsets over which to form generalizations. Thus, at a completely formal level, the third premise of the PotS argument is superfluous: no conceivable input set, consisting of combinations of positive and negative examples, could truly uniquely specify anything. Clearly, the generalization process must be constrained to favor some kinds of subsets a priori. The crux of the argument, then, really resides in the second premise: can the same definition of what makes one generalization better than another be used across domains, or will distinct principles need to be invoked for different spheres of human experience?

The answer to the preceding question will depend heavily on how one believes generalizations are chosen. On a purely deductive account of learning, one might first rule out all generalizations that are contradicted by the data, and select from among those that remain based on some measure
of their complexity, in accordance with a version of Occam’s Razor. Perhaps the hypothesis with the fewest ‘moving parts’ (e.g., category types) should be chosen. Or, perhaps it is best to choose the narrowest available generalization to maintain maximum falsifiability upon encountering further data. Certainly in the latter case, and possibly in the former as well, depending on how one measures moving parts, one is unlikely to prefer an infinitely recursive tree-based grammar over one in which long-distance dependencies are minimized or non-existent, and/or whose generative capacity is finite. Hence, if a deductive learner is to arrive at the correct hypothesis, she would seem to need some a priori guidance beyond mere “simplicity”.

If, however, generalizations are not regarded as either ‘consistent with’ or ‘contradicted by’ the data, but instead are probabilistically supported in a gradient way, the picture becomes significantly more nuanced. In particular, if inference is probabilistic and abductive, in the manner discussed in 1.3.3.1, then it can be shown that, under reasonable assumptions, a structure-dependent grammar is better supported by the evidence than a linear, structure-independent grammar (Perfors, Tenenbaum, & Regier, 2006). If the assumptions underlying this finding are valid, it casts serious doubt on the second premise of the Poverty of the Stimulus argument, thereby undermining the case for a rich base of innate, domain-specific constraints on inference.

The framework of Bayesian probabilistic inference (e.g., Tenenbaum, Griffiths, & Kemp, 2006) provides a principled, domain-general means for selecting among available generalizations, which nonetheless allows domain-sensitive biases to emerge as “overhypotheses”, tying together more specific biases within a domain (Kemp, Perfors, & Tenenbaum, 2007). By structuring the relationships between representations in an explanatory way, highly abstract generalizations can be learned from relatively little information (Goodman, Ullman, & Tenenbaum, 2009), often before the details are worked out. Contrast this with purely “inductive” models such as feedforward connectionist networks, which build up abstract knowledge from specifics, and often either require thousands of training examples to learn even simple, unambiguous generalizations. The ability of hierarchical Bayesian models to learn from few examples by carrying out probabilistic inference over structured
representations is what Goodman refers to as the “blessing of abstraction”, providing a counter-
weight to the “curse of dimensionality”.

1.5 Summary

In this chapter I have laid some of the philosophical groundwork for the study of human inference. I have attempted to characterize forms of natural, cognitive inference by way of analogy to the philosophical formalisms of logical reasoning, as well as the computational formalisms of machine learning. I have argued that employing a structured model of the causal relationships in the environment in the service of inference may confer important advantages over less principled, model-free associative approaches to inference, and that allowing such a model to represent probabilistic relationships allows for much greater flexibility than more rigid, rule-based deductive approaches. Moreover, it is suggested in section 1.4 that some of the key logical arguments in favor of a rich set of innate, domain-specific constraints on learning lose much of their force if inference is probabilistic and model-based.

In the next chapter I review some of the psychological literature on learning in human infants and adults, and focus particularly on the evidence for both probabilistic and model-based inference. I then introduce a particular learning problem, namely that of inferring the existence of abstract, relational ‘rules’. This form of learning has played an important part in discussions of what inference mechanisms are available to children, as well as whether those mechanisms might be designed to apply specifically to some cognitive domains (such as language) and not others. I will argue that, while some sensitivity to domain is undoubtedly present in human inference, these discrepancies can at least in principle be accounted for by higher-order, domain-general inference processes. After reviewing some empirical evidence for higher-order learning in human infants, I will present an account of how the acquisition of a probabilistic, generative model of a domain may be one way for domain-sensitivities in learning to emerge. In the chapters that follow, I will explore this possibility with a set of behavioral experiments, as well as a formal probabilistic model that produces second-
order learning behavior that is parallel to that exhibited by human learners.
CHAPTER 2

A REVIEW OF RELEVANT RECENT EMPIRICAL RESEARCH

In this chapter, I will review some of the empirical and computational research which has shed some light on the questions discussed in Chapter 1. I discuss the rich recent body of work concerning “inductive” statistical learning in various cognitive agents, including human infants and adults, a variety of non-human animals, as well as artificial systems that serve as models of natural learners. Next I review some of the cases in which abductive inference appears to be at play, even in infants and young children. Following this initial discussion of two forms of learning, each grounded in probability, I turn to the debate over the phenomenon of “rule-learning”, and whether it draws on distinct learning mechanisms from those which track statistics such as the algorithms carried out in connectionist networks. Finally, I review a set of findings related to what can be called “metalearning”, or plasticity of the learning system itself. I argue that metalearning can explain some of the findings in the rule-learning literature that have previously been cited as evidence for innately domain-specific learning mechanisms. Moreover, I argue that, in the types of rule-learning situations in question, the type of metalearning needed to explain the empirical results is supported by rational, abductive inference. In summary, while rule-learning appears to be carried out in a domain-sensitive way, the role of domain is neither absolute nor innate. Rather, a strong possibility exists that gradient sensitivity to domain emerges as a result of hierarchical causal inferences.

2.1 Statistical Learning in Infants

The Poverty of the Stimulus argument tacitly assumes that the child performs a kind of “induction by deduction”. That is, given a set of sentences known to be grammatical, the child arrives at a single conclusion about what “grammatical” means for her language, ruling out some hypotheses
that contradict the data, and settling on one of the remaining hypotheses because it is the most natural (whether or not the meaning of “natural” can differ as a function of the input domain). However, since at least the advent of connectionism (Rumelhart & McClelland, 1986) as a serious model class for cognitive architectures, the notion that learners really do arrive at a single, discrete hypothesis has been called into question. In some sense, this amounts to questioning the fifth premise of the PotS argument, namely that children really do possess a representation of a single “correct” grammar — in general, connectionists deny that grammar is explicitly represented at all. If it is not necessary that data lead uniquely to a hypothesis, then the range of learning mechanisms available to a hypothetical domain-general learner is much wider.

Although the connectionist models of the 1980s represented a theoretical formalism for statistical learning, it was not until the mid 1990s that serious empirical investigations of statistical inference in infant learners took place. In a groundbreaking series of studies, Saffran and colleagues showed that human adults, as well as 8-month-old infants, after being familiarized with monotone syllable streams consisting of concatenated pseudowords like “golabu” and “pabiku”, were able to differentiate component sequences from sequences composed of parts of multiple vocabulary items, such as “bupabi” (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Even if the absolute frequency of the items in the familiarization set was controlled for, infants could nonetheless differentiate component sequences from part-sequences (Aslin, Saffran, & Newport, 1998), apparently using some form of conditional probability to differentiate the two types of sequences.

The use of some form of statistical information in inferring component units from an ambiguous composite appears to be widespread across cognitive domains, both in auditory and visual modalities. Discrimination of component and part-sequences has been found with sequences of tones (Saffran, Johnson, Aslin, & Newport, 1999), and an analogous capacity appears to be present for “shape scenes” as well (Fiser & Aslin, 2001, 2002). This capacity does not appear to be unique to humans, either: Hauser, Newport, and Aslin (2001) found that cotton-top tamarins successfully extracted component sequences from continuous streams of speech in much the same way as human
In addition to segmenting ambiguous displays into component objects, statistical learning appears to play a role in clustering tokens into categories. Maye, Werker, and Gerken (2002) found that infants exposed to a continuum of syllable tokens varying along the voice-onset time (VOT) dimension treated the continuum as containing two categories if the VOT distribution was bimodal, but only one category when a unimodal distribution was used. This type of distributional learning appears to induce both “lumping” and “splitting” behavior, with previously distinct tokens becoming merged after unimodal exposure, as well as two tokens previously treated as the same being differentiated following bimodal exposure (Maye, Weiss, & Aslin, 2008).

2.2 Abductive Reasoning and Explaining Away

The project of inferring the most probable causal theory from observable data is the fundamental task facing any scientist. We collect information about variables that we can measure directly, and select from among hypothesized causes based both on prior plausibility (i.e., based on what is known about the domain more generally) and on compatibility with the data (the observed pattern should not be too surprising if the hypothesis is true). In many ways, the cognitive task of the infant, newly cast into the world, is of the same kind: She does not have direct access to the kinds of representations of her environment that are most important for guiding behavior — the emotional and mental states of those around her, the meaning being expressed by the language she hears, even the fact that her mother’s face is nearby. Instead, she must infer these properties from simpler features — the shape of a mouth, the presence and arrangement of certain words. And even these features must be inferred from raw sensory data — patterns of light, sound, smell and touch. Broadly, perception, language learning, casual causal reasoning, and formal science are the same kind of problem. Moreover, they are all carried out by the same squishy mass. But do the

\[\text{It should be noted that the first author of this paper is on leave from his position at Harvard after an investigation confirmed multiple instances of scientific misconduct, and several of his published works have been retracted (Miller, 2010). Although the paper cited here has not been retracted, its findings should be regarded with some skepticism.}\]
similarities end there, or is there a deeper connection?

In recent decades, developmental psychologists have formalized this analogy in greater detail (see Gopnik, 2000, for a review), even turning it around to suggest that, as Gopnik memorably puts it, “It is not that children are little scientists but that scientists are big children.” More recently, developmental cognitive scientists have begun to synthesize this “theory theory” of development with the notion that the human brain is a powerful statistical computer. The result is the emerging framework of Bayesian probabilistic models of cognition (e.g. Tenenbaum et al., 2006).

One notable piece of evidence that infants employ probabilistic causal theories comes from Xu and Garcia (2008). In that study, 8-month-old infants were shown displays of red and white balls being drawn from a hidden container. After seeing samples containing mostly red balls, the infants were surprised if the larger container was later shown to contain more white balls than red (and vice versa). Conversely, if the larger container was shown first but the samples were hidden, the infants were surprised when the sample distribution deviated markedly from the population. Xu and Garcia argued that these findings showed that the infants possessed a model of a random sampling procedure, with a representation of the likelihood of obtaining samples with particular characteristics. In a subsequent study, Xu and Denison (2009) replicated these findings with 11-month-olds, but found that the pattern of surprise changed if the samples were drawn not by a random process, but by an experimenter with previously demonstrated preferences for a certain color ball. For example, infants who had previously seen the experimenter selectively drawing red balls were surprised if a sample drawn by the same experimenter contained mostly white balls, even if the population had a similar distribution. If, however, the experimenter wore a blindfold and hence could not see what color she was sampling, the pattern of surprise resembled that when no preference had been expressed.

In a related study, Kushnir, Xu, and Wellman (2010) found that 20-month-olds’ willingness to attribute a preference for a particular type of toy to an agent was sensitive to the prevalence of that toy type in a collection. If the agent chose a toy that is rare in the population, infants were more
likely to conclude that their choice was based on a preference than if the toy was more common. Gweon, Tenenbaum, and Schulz (in press) present a Bayesian model that reproduces these results, with a parameter governing whether sampling occurs randomly from an entire population (“weak sampling”), or whether samples are drawn because they possess a particular property (“strong sampling”).

There is a rich literature emerging in developmental psychology concerning the ability of infants and young children to simulate the minds of others — attributing motivations, beliefs and intentions to other agents, and interpreting their actions through the lens of these latent mental states. This tendency is not usually discussed in the context of abductive reasoning, but it is a very natural example: Simulation requires a theory which makes predictions about the observable behaviors that will result when a particular mental state is present (indeed, this sort of simulation is usually referred to as “theory of mind”).

In a landmark study, Gergely, Nádasdy, Csibra, and Bíró (1995) found that 12-month-old infants appeared to infer that the action of a simple geometric shape was goal-directed upon watching it follow a path around an obstacle. The infants subsequently found a display in which the shape followed the same trajectory in the absence of the obstacle more surprising than one in which the shape instead took a more direct route. The former display was perceptually more similar, but was nonetheless treated as more different by the infants. In a study by Gergely, Bekkering, and Király (2002) 14-month-old infants observed an experimenter employ an unfamiliar head action which resulted in the illumination of a light-box. In one condition the experimenter’s hands were free; in another they were being used to hold a blanket. Infants in the hands-free demonstration condition were significantly more likely to imitate the head action than those in the hands-encumbered condition. Consider the implicit reasoning process required to produce this difference. When the experimenter uses her head even though she could have used her hands, this is taken as evidence that the head action has a special function; however, if the experimenter’s hands are occupied, using the head may simply reflect an effort to make contact with the light-box through whatever
means are available. Range, Viranyi, and Huber (2007) found similar behavior in dogs: dogs who observed a conspecific pulling a lever with its paw would imitate the action exactly when the other dog used its paw when it could have used its mouth instead; but if the other dog had a ball in its mouth, the observer would instead pull the lever with its mouth (a more natural action for a dog).

The abductive inference made by the infants in Gergely et al. (2002) and Range et al. (2007) is an instance of explaining away, as it is described in section 1.3.3.2. There are at least two hypotheses available to explain the observed head action: On the first hypothesis, any contact will illuminate the light-box, whereas under the second hypothesis, contact with the forehead is required. Under hypothesis 2, the head action is likely to occur regardless of the state of the experimenter’s hands, but under hypothesis 1, the action is only likely when a hand action is unavailable. Thus, in the hands-free condition, the second hypothesis is more strongly supported by the evidence, whereas the occupied hands explain away the evidence for the narrower mechanism, and the hypothesis which is more plausible \textit{a priori} will be preferred.

It may be that the domain of mental states, goals and intentions is a special case in which employing a truly causal model is especially useful. Moreover, it is possible that mental simulation of actions resulting from mental states is easier or more natural than simulating other domains; after all, the processes in question are taking place in a brain already. In other domains, learners may fall back on comparatively simpler associative learning mechanisms, inductively learning about relationships among environmental variables without modeling causal relationships. However, a number of studies suggest that abductive inference plays a role in non-mental domains as well.

Sobel, Tenenbaum, and Gopnik (2004) used a ‘blicket detector’ paradigm (Gopnik, Sobel, Schulz, & Glymour, 2001) to show that preschool-age children engaged in causal reasoning that could not be explained by associative mechanisms. In this study, children were shown a novel machine which they were told lit up when a ‘blicket’ was placed on it. They were then shown that the machine lit up when two objects, A and B, were placed on the machine simultaneously. The children then saw object A placed on the machine alone. In the \textit{one cause} condition, the machine did not light up; in
the *backwards blocking* condition, it did. Children in the one-cause condition were confident that object B would light up the detector, but those in the backwards blocking condition were unsure. If associative strength were driving children’s judgments about blicket status, there should have been no difference between the two conditions, since in both cases the machine lit up in the presence of B every time. However, if the reasoning process is explanatory in nature, then the pattern of results exhibited is exactly what should be predicted. Indeed, the situation in the backward blocking condition is formally identical to the rain and sprinkler scenario described in section 1.3.3.2: the initial evidence is consistent with two hypotheses which are unrelated *a priori*: (1) A is a blicket, and (2) B is a blicket. Then, new evidence is observed — A makes the detector light up — which is consistent with one hypothesis but has no bearing on the second. This new evidence explains away the original evidence in favor of the second hypothesis.

Sobel and Kirkham (2006) adapted the blicket-detector paradigm to 8-month-old infants. Using an anticipatory eye-tracking measure, they presented infants with a display involving sequences composed of four events. Events A and B were initially presented simultaneously, and were always followed by event C (analogous to the blicket detector lighting up). Then, in the *indirect screening-off* condition, event A was presented alone and was always followed by event D (analogous to the blicket detector failing to light up). In the *backwards blocking* condition, A was always followed by C. At test, B was presented alone. As with the preschoolers, infants in the indirect screening-off condition were more likely to engage in anticipatory looks toward the location where event C had occurred, suggesting that A being followed by C explained away the evidence that B would be followed by C.

### 2.3 Learning Abstract Rules

It has become fairly clear that humans (as well as other animals) possess powerful and domain-general abilities to track statistical information in their environment, which is used in segmentation (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Aslin et al., 1998; Saffran et al.,
However, it remains to be seen to what extent other, domain-specific mechanisms are involved in learning about complex systems such as language. Indeed, it is perfectly plausible that, while statistical learning is useful for sorting through the levels of linguistic structure that are in some sense “closer to the signal” (such as phonology and the lexicon), qualitatively different, and more specialized, processes are needed to infer the “deeper” kinds of structure needed for syntax. In more recent writings, Chomsky and his colleagues have indeed suggested that the hypothesized language faculty may really only contain syntactic information (Chomsky, 1995; Fitch & Hauser, 2004; Fitch, Hauser, & Chomsky, 2005).

Gómez and Gerken (1999) found evidence for grammar acquisition in 12-month-olds familiarized with sequences of nonsense syllables generated with a Finite State Grammar (FSG). The infants in their experiments generalized not only to new sequences generated by the FSG, but could also distinguish grammatical from ungrammatical sequences when each syllable in the training set was replaced by a novel syllable.

Marcus, Vijayan, Rao, and Vishton (1999) found that 7-month-old human infants could extract what they termed an “algebraic rule” from sequences of syllables. After familiarization with sequences like “ga ti ti” and “le di di”, infants distinguished novel sequences like “ba po po” from sequences whose individual syllables were equally novel (like “po po ba”), but which violated the abstract repetition rule that bound together the familiarization sequences.

The key difference between the Gómez and Gerken (1999) and Marcus et al. (1999) experiments, on the one hand, and earlier statistical learning experiments such as those of Saffran, Aslin, and Newport (1996) and Maye et al. (2002), on the other, is the ability of infants to apply a learned regularity that is sufficiently abstract as to determine the grammatical status of sequences which share none of the individual elements from the training set. It is this abstraction that led Marcus et al. (1999) to claim that infants might have access to a rule-learning mechanism which is qualitatively distinct from the statistical learning apparatus employed in segmentation and clustering.
tasks, and that this abstract learning ability is beyond the capacity of many existing statistical learning formalisms such as Simple Recurrent Networks (Elman, 1990), and other connectionist architectures. This claim has been hotly debated (Seidenberg & Elman, 1999a; Marcus, 1999a; Altmann & Dienes, 1999; Marcus, 1999c; Seidenberg & Elman, 1999b; Marcus, 1999b; Christiansen & Curtin, 1999; Marcus, 1999d), with its critics pointing out that connectionist networks have internal layers of representation that need not map transparently to surface features.

The debate over the existence of a rule-learning ability that relies on a cognitive mechanism that is not statistical in nature is further complicated by a set of results obtained by Marcus, Fernandes, and Johnson (2007). They find that the ability of 7-month-old infants to learn AAB-style rules of the sort exemplified by “le le di” and “po po ga” is diminished or non-existent when the A and B elements in question are non-linguistic sounds, such as musical tones, instead of syllables. While the segmentation abilities of the infants in Saffran, Aslin, and Newport (1996) appear to be highly domain-general (Saffran et al., 1999; Fiser & Aslin, 2002), the failure of more abstract rule-learning to generalize to tones was interpreted by Marcus et al. as evidence for a distinct mechanism, perhaps specially suited to linguistic processing. This claim is empirically undermined, however, by Saffran, Pollack, Seibel, and Shkolnik (2007), (Johnson et al., 2009) and Murphy, Mondragon, and Murphy (2008), who find evidence for AAB-style rule-learning in other modalities (visual objects and shapes) and species (rats). Moreover, Dawson and Gerken (2009) find that infants three months younger than those in Marcus et al. are capable of extracting AAB and ABA regularities over the very same kinds of elements (musical tones) that led to failures of learning in that study. This last finding is particularly perplexing, as it apparently represents a loss in a learning ability during the course of development.

2.4 Adaptive Perception and Learning

There is a long history of research documenting some form of adaptive loss of sensitivity in infants during the first year of life. A classic example comes from Werker and Tees (1984), who found
that infants seem to lose sensitivity to nonnative phonetic contrasts some time between 8 and 10 months of age. (Bosch & Sebastián-Gallés, 2003) extended this pattern to bilingual infants, finding a U-shaped sensitivity curve in infants exposed to two languages, only one of which employed the phonetic contrast in question. Stager and Werker (1997) found that the degree of phonetic sensitivity varies as a function of context, with finer distinctions made in a syllable discrimination task than in a word-learning task. The distributional learning seen in Maye et al. (2002) provides a plausible mechanism for these shifts.

In the domain of music, there appears to be a shift in the salience of different pitch relations as listeners acquire experience with melodies in their culture. When tone streams, analogous to the syllable streams from Saffran, Aslin, and Newport (1996), are segmentable either by absolute pitch or relative pitch (i.e., intervals), 8-month-old infants rely on the former, whereas adults rely on the latter (Saffran & Griepentrog, 2001). This shift is adaptive, since melodic identity is determined primarily by the pattern of intervals between pitches, and not the identity of individual pitches. Similarly, the ability to detect melodic changes that do not affect the diatonic role of a pitch declines some time after 8-months of age, whereas diatonic changes with a smaller acoustic magnitude remain salient (Trainor & Trehub, 1992). Lynch and Eilers (1992) report evidence that 12-month-old infants are better able to detect mistunings in melodies constructed using a Western major scale than in either of two other, less familiar scale structures. Even by six months, infants in Lynch and Eilers’s study were better at detecting mistunings when the scale structure contained an underlying semitone structure. Hannon and Trehub (2005a) found cultural differences in the perception of rhythmic contrasts, with North American adults failing to detect a rhythmic contrast that is used in Eastern but not Western European music, while 6-month-old infants and Bulgarian adults succeeded. This rhythmic assimilation appears to be in place by 12 months of age (Hannon & Trehub, 2005b), although brief exposure in the laboratory to music in which the contrast is relevant is enough to restore sensitivity.

Adaptive changes in sensitivity appear to take place not only in perception, but in more abstract
learning tasks as well. Gerken and Bollt (2008) found that, while 7.5-month-old infants could learn an artificial stress rule based either on syllable heaviness (a property which contributes to stress in many natural languages) or on the presence of a particular syllable-initial consonant (a feature which is not commonly associated with stress), slightly older infants could only learn the linguistically natural rule. This result suggests that by 9 months of age, infants acquire some higher-order principles pertaining to which features are likely to be relevant for stress.

2.5 Possible Adaptations in Rule Learning

It is an open question why infants appear to lose sensitivity to repetition rules with tone stimuli between 4 and 8 months of age, as evidenced in the experiments of Marcus et al. (2007) and Dawson and Gerken (2009). Given the younger infants’ success, as well as the success of older infants with visual stimuli (Saffran et al., 2007; Johnson et al., 2009), and of rats with auditory stimuli (Murphy et al., 2008), it seems likely that the capacity to learn abstract repetition rules begins as domain-general. While it is possible that an experience-invariant developmental process restricts this ability to a subset of the domains studied in human infants, a possibility that is both more interesting and more plausible, in my view, is that infants are adapting to the structure of their environment in a domain-sensitive way.

There are several ways in which this might occur. One is that the relevant learning is a result not of experience with music, but of experience with language. All of the infants studied in both Marcus et al. (2007) and Dawson and Gerken (2009) were living in an English-speaking environment. In the English language, pitch is commonly used to convey affective information (Trainor, Austin, & Desjardins, 2000), and so it is possible that through linguistic exposure, infants have learned by 7-months that pitch is less likely to be used to convey abstract, structural information than to convey emotional content, which is in some sense “analog”. However, there is evidence that prosody conveys syntactic information as well (see, e.g., Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989; Fisher & Tokura, 1996), and so it is far from obvious that disregarding potential abstract
structure that depends on pitch relations would be adaptive. Nonetheless, it would be useful to explore whether infants in language environments where pitch is more explicitly structural (say, Cantonese) have an easier time learning abstract tone rules.

A second class of domain-sensitive adaptation relies on infants’ experience with music. It is possible that changes in the way infants represent or attend to sequences of tones could increase the difficulty of extracting an abstract rule from such sequences. In order to detect the identity relationship that exists between two particular elements in a sequence, it is necessary to encode tone sequences hierarchically, treating both individual pitches (or chords) and three-tone (or chord) sequences as constituent units. If by seven months of age, infants have transitioned to a holistic representation of melodies, no longer maintaining a strong representation of individual pitches, it would be difficult for them to succeed in learning about the internal “syntax” across a set of AAB sequences.

A related hypothesis is that 7-month-olds place a greater focus on pitch contour 4-month-olds do, and hence may have a difficult time learning a rule that requires them to collapse across multiple contours. However, it appears that contour is quite salient from a young age: Chang and Trehub (1977) find that infants are robustly sensitive to contour at least by 5 months, and in an unpublished experiment from our lab, 4-month-olds were able to learn a contour rule that required them to categorize U-shaped three-tone melodies as distinct from inverted U-shaped melodies. Hence, it is not clear that collapsing across contours should not be as difficult for younger infants as it is for older ones.

In Dawson and Gerken (2009), we suggested a form of adaptation which, like the preceding factors, depends on infants developing new representations of music, but which makes a critically distinct hypothesis about the nature of learning. What we proposed is that, rather than simply attempting to form a generalization that captured some regularity in their input, infants form hypotheses about the underlying processes that generated the input in the first place. In other words, generalization occurs in the name of explanation, rather than merely prediction.
2.5.1 The Principle of Pitch Proximity, or “Smoothness”, in Music

It has been formally documented for over 80 years that natural melodies across many cultures obey the principle of “pitch proximity” (e.g., Watt, 1924; Ortmann, 1926; Dowling, 1967), whereby small intervals consistently outnumber large ones. In other words, tones that occur nearby in time also tend to be close together in pitch.

Both production and perceptual factors have been identified as potential sources for this property. On the production side, producing large intervals on nearly any instrument, including the human voice and with the sole exception of some electronic instruments, requires more physical movement by the player. Hence, a preference for small intervals is in keeping with the Principle of Least Effort (Zipf, 1949). For listeners, pitch proximity serves as an acoustic cue to sequential grouping (Bregman & Campbell, 1971; Bregman & Pinker, 1978). When two alternating tones are played at sufficiently high speeds, they are perceived as emanating from two distinct sources if their frequency ratio is high enough. The farther apart in pitch the two tones are, the slower the stream must be played in order to produce a percept of a single stream.

In (Dawson, 2007), I found that if the size of the interval between successive pitches is encoded using steps on the diatonic scale (as opposed to an “objective” acoustic measure of the distance between two pitches), then the overall distribution of intervals in a corpus of children’s folk songs can be fit quite well with a Normal distribution. In particular, the probability of a repeated tone occurring at any given point is higher than the probability of any other interval, but not significantly more frequent than expected given a model of interval generation in which the probability of a particular pitch falls off gradually with its distance from the preceding pitch. As such, a learner who utilizes a mental model of music that contains this “Smooth” property should find the occurrence of repetitions less surprising than would a learner who has not internalized this property of music. If the occurrence of repetitions is largely predictable as a result of a general bias toward small intervals, then there should be little reason to posit the existence of a special rule requiring repetitions to
occur at a particular position in a melody.

Inferring the existence of hidden causes of one’s input (e.g. “repetitions are frequent because intervals tend to be small”, or “repetitions are frequent because sequences are generated using an abstract rule”) is a form of abduction. As such, it should naturally be produced by an abductive learner. Indeed, pitch proximity is one of three principles (alongside tessitura and key profile) incorporated in a recent generative model of melody perception (Temperley, 2008).

2.6 Empirical Research Reported in This Dissertation

The principal question addressed by the empirical research reported here is whether, indeed, human learners (both infant and adult) will exhibit “explaining away” behavior of the sort described in Dawson and Gerken (2009). Specifically, if they are presented with a melodic environment in which individual intervals between successive pitches have an overall tendency to be small in magnitude, will the increased repetition rate that is expected as a consequence of this fact inhibit inferences about an explicit repetition rule? Experiment 1 in Chapter 3 asks this question with adult learners by manipulating the interval distribution to which learners are exposed immediately prior to learning a repetition rule of the form studied in Dawson and Gerken (2009). In Chapter 4, a simple probabilistic model in the form of a Bayes Net examines the “ideal” inferences that are licensed when two conditions are satisfied: (1) full information about the range of processes used to generate the data is available, and (2) beliefs about the environment are updated in an optimal Bayesian fashion. In Chapter 5, Experiment 2 adapts the same experimental paradigm to 4- and 7.5-month-old infants, the same ages studied in the 2009 paper.

In Chapter 6, Experiment 3 addresses a possible alternative interpretation for the first two experiments, namely that the presence of Smooth melodies affects participants’ engagement in a way that interferes with learning generally, by examining the effects of the same context manipulation on the learning of a grammar that is more consistent with the nature of a Smooth environment than with the other two context environments used. If the interval distribution present in the context
is really being used abductively as evidence about the abstract nature of the environment, then the Smooth environment should produce the opposite effect here, or at the very least, should not impair performance in the same way as in the previous experiments.

The final chapter provides a general discussion of the three experiments and model, and in it I suggest some future research to address some ambiguities in the present data, as well as to further explore the extent to which explanatory theory-building provides a good model of humans’ ability to learn about the abstract structures present in domains such as language and music, among others.
CHAPTER 3

EXPERIMENT 1

3.1 Introduction

In music, there are at least two ways to represent tone events, and the relationships among pitch values. In one representation, individual tones function as though they were discrete “words”. In this mode of representation, similarity in surface form need not imply similarity in functional significance. This property of arbitrariness of the sign (Saussure, 1916) is generally considered to be one of the defining characteristics of truly symbolic systems such as language (though see Monaghan & Christiansen, 2006; Monaghan, Christiansen, & Chater, 2007, for evidence of partial form-meaning systematicity in language at the level of syntactic categories). An alternative representation of tone events is not as abstract, symbolic types, but as points on a pitch continuum. These two modes of representation need not be mutually exclusive — indeed, natural music has properties of both discrete and continuous systems. Consecutive points on the scale, such as the tonic and the leading tone, can play very different roles in the Western tonal system, reflecting a separation between “form” and “function”. At the same time, most melodies tend to move “smoothly”, with relatively few large leaps between temporally consecutive tones (cf. the discussion on “pitch proximity” in Section 2.5.1), reflecting the pitch continuum underlying the functional system of a scale structure.

Repetition can be seen as playing two different roles in music: one discrete and one continuous. On the one hand, it constitutes an identity, or “sameness” relation between two tones. It is also an interval of magnitude zero. If one assumes that any tone is equally likely at any point (the tone distribution is uniform), hearing every melody begin with two repeated notes would be quite surprising, and evidence for an identity interpretation would be strong. If, however, one knows that
tones nearby in time also tend to be nearby in pitch (melodies are usually “smooth”), a high rate of repetition (qua interval of distance zero) should be expected as a natural consequence of this “Smoothness”, and it should take more evidence to conclude that repetition is special.

In Dawson and Gerken (2009), we hypothesized that, after 7 months of experience with naturally Smooth Western music, infants may have internalized that general property of melodies, leading to a greater threshold of sensitivity to repetition as anything other than merely a small interval. In turn, this could lead to greater difficulty learning a rule like AAB\(^1\). In that paper, however, the evidence for such an interpretation was quite indirect. Here, I attempt to collect more direct evidence for the effect of Smoothness on sensitivity to repetition, by directly manipulating listeners’ musical context immediately prior to learning a repetition-based rule. In this first experiment, I investigate this effect with adult listeners. In Experiment 2, the same procedure is adapted to 4- and 7.5-month-old infants.

Since small intervals are not surprising in a Smooth environment, a learner modeling this tendency should not treat frequent repetitions as indicative of additional structure. Learners previously exposed to the Smooth environment should less readily infer the existence of a repetition rule than those familiarized with non-Smooth melodies. On this account, it is not merely the high rate of repetition in the Smooth environment that leads to discounting of the evidence for a repetition rule; rather, it is that a high rate of repetition is an incidental consequence of the Smoothness property. If the environment contains a large number of repetitions in the absence of a Smoothness Constraint (that is, unexplained repetitions), no discounting should occur.

The central prediction is that a repetition rule will be more difficult to learn in a Smooth melodic environment. To test this, participants are placed in one of three melodic contexts. In the Uniform condition, every tone is equally likely at any point. In the Smooth condition, small intervals are more common than large intervals. In the Repetition condition, repetition alone is more frequent than other intervals. The latter two groups are subdivided into high repetition (Low Variance) and

\(^1\)In a Smooth context, non-adjacent repetitions, such as those found in an ABA grammar, should be frequent as well, and indeed, intervals over a ‘lag’ of 2 notes can be fit by a Normal distribution as well (Dawson, 2007).
low repetition (High Variance) conditions, with the absolute repetition rate matched between each Smooth group and its Repetition counterpart. This latter manipulation is intended to examine whether the quantity of repetitions has an influence on grammar-learning over and above that of the qualitative shape of the distribution. Plausibly, hearing more repetitions in the context could desensitize learners, leading to a decreased grammar-learning ability; alternatively, the presence of unexplained repetitions could drive learners to devote more attention to that feature. Indeed, both of these could occur simultaneously, leading to the appearance of little to no effect of repetition rate. On the explaining away account, however, any effect of repetition rate should be smaller than the effect of qualitative distribution type.

Following exposure to these contexts, participants perform a grammar-induction task where. Half of participants learn a repetition-initial (AABCD) grammar, and half learn a repetition-final (DCBAA) grammar. If learners model the overall interval distribution, the Smooth context should emphasize the zero-magnitude interval interpretation of repetition, which is expected to be frequent, instead of the grammatical feature representation. Learners in this context should exhibit decreased sensitivity to positional repetition, as well as decreased grammar-learning performance. No explaining-away should occur in the Repetition groups, since no broader property accounts for the high rate of repetition. If, on the other hand, learners are not attempting to explain the distribution of intervals in the environment, but are merely attempting to identify distinctive properties of the grammatical training sequences, then a different pattern of performance should obtain. The dominant predictor of performance should be repetition rate, with the groups exposed to the greatest quantity of repetition (namely the Low Variance Repetition and Smooth groups) faring best, followed by the High Variance Repetition and Smooth groups, with the Uniform group exhibiting the poorest performance.
3.2 Methods

3.2.1 Participants

One hundred and twenty University of Arizona undergraduates participated in the study for course credit. An additional eighteen participated but were excluded from analysis due to their failure to score above chance on a melodic-discrimination screening task.

3.2.2 Materials and Procedures

The experiment consists of a Context phase and a Grammar-Learning phase. The latter contains four blocks, each with a training component and a test component. All “sentences” consist of five tones generated using the FM Synthesizer in the MIDI Toolbox for MATLAB (Eerola & Toiviainen, 2004), which produces a horn-like sound. The first four notes are 250 msec each, with 50 msec gaps after each one. The last note is 500 msec. In music terms, the melodies contain four eighth notes followed by a quarter note, played at 200 beats per minute.
Figure 3.2: Interval Distributions in the Context Phase. Here, 0 is a repetition, +1 is a step to the next-highest note, etc. The distributions are combined across all pitches. The truncation of the pitch range results in more small intervals across all conditions. If the interval distributions were separated by preceding pitch, those for the Uniform and Repetition conditions would each be flat, except for the peak at 0 in the Repetition case.

3.2.2.1 Procedures: Context Phase

The Context phase consists of two blocks of 100 sentences, in random order. Ten are probe sentences, after which either the same sentence is repeated or one of the other ten probe sentences is played. On the probe trials, participants have 3 seconds to press the 1 or 0 key on the keyboard to register “same” or “different” sentence pairs. The absence of a response is coded as incorrect. Each block lasts about five minutes. Data from participants who did not perform above chance on this discrimination task (15 or more out of 20 correct) was discarded, as these participants presumably either could not distinguish differences among melodies, or were not attempting to succeed. During context exposure, all participants see a group of eight cartoon aliens (Folstein, Van Petten, & Rose, 2007). Half are “star-chested” and half are “brick-chested” (Fig. 3.1).
3.2.2.2 Materials: Context Phase

Participants are assigned to one of three context conditions: Uniform \((n = 24)\), Smooth \((n = 48)\) or Repetition \((n = 48)\). The Smooth and Repetition conditions are further divided into High Variance (HV) and Low Variance (LV) sub-conditions. In all cases, context melodies are drawn from a vocabulary of six tones: A3, A♯3, C♯4, E4, G4 and G♯4 (MIDI values 57, 58, 61, 64, 67 and 68).

In the Uniform condition, each tone is equally likely and independent of the last. As such, the probability of a repetition at any given point is 1/6 (in the 200 generated melodies, the empirical rate was 18.1%). The resulting distribution of intervals is shown in Figure 3.2a.

In the Smooth condition, melodies are generated as follows. The first tone is chosen from a uniform distribution over the six tones. For each subsequent tone, a sample is generated from a normal distribution, truncated between 0.5 and 6.5. The mean of the distribution is an integer corresponding to the previous tone (the lowest tone is 1; the highest tone 6). The standard deviation is 2 in the HV condition and 1.2 in the LV condition. The sampled value is rounded to the nearest integer to generate the next tone. The resulting “discretized Normal” distribution reflects the interval distribution of typical folk music (Dawson, 2007). The rate of repetition across the 200 melodies is 39.3% of all intervals in the LV condition (Fig. 3.2b), and 26.3% in the HV condition (Fig. 3.2c).

The Repetition conditions control for the actual rate of repetition, while removing the overall Smoothness constraint. Here, the HV and LV conditions (Fig. 3.2d-e) are matched to their Smooth counterparts for the number of repetitions, but unlike in the Smooth cases, the remaining notes are equiprobable. Here, the high rate of repetition cannot be explained by a general bias for small intervals; instead, a learner modeling the tone distribution must encode repetitions separately to achieve a good fit.
3.2.2.3 Procedures: Grammar-Learning Phase

After the context phase, participants move on to the grammar-learning phase. They are asked to detect “spies” attempting to infiltrate the “Qixian” colony, and are told that they can distinguish Qixians from spies by the grammaticality or ungrammaticality of their speech.

In each training block, participants hear thirty “grammatical” sentences in random order while an image of four star-chested aliens is displayed (Fig. 3.3).

After each training block, participants hear twenty-four test sentences, half grammatical. After each sentence, participants make a continuous grammaticality judgment by clicking on a line (Fig. 3.3b), where the left pole represents “definitely grammatical”, the right pole represents “definitely ungrammatical”, and every gradient response in between is possible. There is no time limit. The computer records a binary response, based on whether the participant clicks left or right of center, and a continuous “discrimination score” calculated by subtracting from 1 the proportion of the line lying between the response and the correct pole. Participants experience four training-test cycles on the same grammar.
3.2.2.4 Materials: Grammar-Learning Phase

The “Qixian” and “spy” sentences are again five tones in length. Each participant is trained using one of two five-tone vocabularies. The first (V1) contains the tones A3, C4, D4♯, F4♯ and G4 (MIDI 57, 60, 63, 66 and 67); the second (V2) contains the tones A♯3, B3, D4, F4 and G♯4 (MIDI 58, 59, 62, 65 and 68). Each set shares two tones with the context vocabulary.

For half of participants, the “grammatical” sentences follow an AABCD pattern (with a repetition at the beginning and nowhere else), while the “ungrammatical” sentences have a DCBAA pattern. For the other half of participants, the labels are reversed.

Of the 120 sentences possible in each grammar, 60 are used as training items, and 24 as test items. The chosen items were balanced for pitch contour: \( \frac{1}{4} \) in each section had a rising segment followed by a falling segment (in addition to the repetition), \( \frac{1}{4} \) had the reverse; \( \frac{1}{4} \) had a rise-fall-rise pattern and \( \frac{1}{4} \) a fall-rise-fall pattern. A schematic representation of possible sentences in each
of the two grammars is depicted in Figure 3.4

Thirty training sentences are used in the first two learning blocks; the other thirty in the last two blocks. On odd-numbered test blocks, participants are tested with items from the training vocabulary; on even blocks they hear items from the opposite vocabulary. Both vocabularies were used to test whether the context manipulation has an effect on the level of abstraction at which participants learn the grammar. The training vocabulary always comes first, as the vocabulary switch could provide a clue to the nature of the grammar (i.e., that it was vocabulary-independent), and if the new vocabulary came first, participants could not demonstrate mastery independent of this “hint”.

3.3 Results

The central question is whether prior exposure to the Smooth distribution will impair detection of the repetition rule. If so, this will suggest that learners model the full interval distribution, which (partially) explains away the training repetitions. The key comparison is therefore between the Smooth groups and the non-Smooth groups.

A secondary question is what effect the overall rate of repetitions, independent of the presence of a Smoothness constraint, has on learning a repetition rule. If broader structure is irrelevant, and learners are influenced only by the amount of repetition they are exposed to, there are two possibilities: if the presence of background repetition has a desensitizing effect, then the Uniform group should outperform the High Variance Repetition group, which in turn should outperform the Low Variance Repetition group. Similarly, the High Variance Smooth group should outperform the Low Variance Smooth group. If background repetitions serve to highlight identity relationships, the reverse rankings should obtain. If qualitative structure is primary, however, any effect of Repetition Rate and/or Variance should be subordinate to the shape of the interval distribution.

Pilot data revealed that many participants performed near ceiling at discriminating grammatical and ungrammatical sentences, while another large set performed at chance overall. For many of
these, presumably only a fairly strong manipulation would observably shift performance. As such, the particular values of the scores received by these participants are mostly uninformative, and contribute noise that could obscure effects of the manipulations.

To address this issue, participants were separated by quartiles within each context condition based on their combined number of correct responses throughout the four test blocks, and two sets of analyses were conducted. The first used all of the data; the second discarded the highest- and lowest-performing quarters in each condition, thereby greatly reducing the proportion of participants performing either at floor or ceiling. When “floor” is defined as producing fewer than 57 correct binary responses out of 96 (the one-tailed $p < 0.05$ cutoff under coin-flip guessing), and “ceiling” is defined as 88 or more correct (i.e., the same distance from 100% as floor is from 50%), then of the 60 participants in the trimmed sample, only 9 were still at floor, and 7 at ceiling. Of the 30 participants excluded for low performance, all but 2 were at floor, and of the 30 excluded for high performance, all but 4 were at ceiling.

Both the binary (correct vs. incorrect) and continuous (confidence-weighted) responses were analyzed. The results from the binary scores are reported first.

### 3.3.1 Number Correct

#### 3.3.1.1 Full Sample

The total number of correct responses was computed for each participant at each block and these scores were entered into an Analysis of Covariance (ANCOVA) with between-subjects factor Context Condition (five levels: Uniform, Smooth (High Variance), Smooth (Low Variance), Repetition (High Variance) and Repetition (Low Variance)), within-subjects factor Block (1 through 4), and a covariate derived from the number of correct “same-different” responses to the probe trials in the Context phase. The raw same-different scores (out of 20) were converted to z-scores within each Context condition.

Three orthogonal planned comparisons among the five Context conditions were of interest.
Figure 3.5: Mean Percent Correct by Block and Context. Error bars denote ±1 SE.

The comparison of greatest interest is between the two Smooth groups on the one hand and the other three groups on the other. This contrast is henceforth the “Smoothness contrast”. The next contrast (“Repetition Rate”) examines the effect of repetition rate among the non-Smooth contexts. Contrast coefficients for Repetition Rate were spaced according to the raw number of repetitions in each of the three conditions (18.1% of the total number of intervals in the Uniform condition, 26.3% in the High Variance Repetition group, and 39.3% in the Low Variance Repetition group). The final comparison (the “Smooth Variance contrast”) contrasts the two Smooth groups.

For the Block factor, three orthogonal contrasts, based on a polynomial function of the block numbers, were used. The linear term measures constant improvement in performance over time, the quadratic term measures constant changes in the rate of improvement, and the cubic term in this case measures differences in improvements after odd vs. even blocks. Since odd test blocks used the same tone vocabulary as the training blocks, whereas even test melodies used entirely different
tones, this last contrast is potentially of interest.

The “same-different” covariate had a significant positive relationship with correct responses ($F(1, 114) = 12.80, p < 0.001$). This would seem to validate the intuition that low-performance on the same-different task reflects a generally poor ability to encode tone patterns, and/or a lack of effort on the part of participants. This serves as validation for excluding especially poor same-different performers from analysis.

The Context × Block interaction was nonsignificant ($F(12, 345) = 0.46, p = 0.93$). The linear Block term was significant ($F(1, 345) = 68.86, p < 10^{-14}$), with an increase in performance across blocks. The quadratic term was marginally significant ($F(1, 345) = 2.80, p < 0.1$), with performance leveling off overall in the later blocks. The cubic term was nonsignificant ($F(1, 345) = 2.43, p = 0.12$).

Most important, the contrast between the Smooth Context group and the other groups was significant ($F(1, 114) = 6.72, p < 0.02$), with the Smooth group exhibiting lower performance overall. The Repetition Rate contrast was not significant, however ($F(1, 114) = 0.95, p = 0.33$), nor was the Smooth Variance contrast ($F(1, 114) = 0.73, p = 0.39$). Means and standard errors for each distribution group at each block (collapsing the High and Low Variance groups) are displayed in Figure 3.5a.

### 3.3.1.2 Trimmed Sample

The above analyses were repeated using only those participants performing between the first and third quartiles of each context group, as determined by total number of correct responses collapsing across blocks.

Participants’ total correct responses in each block were entered into an ANCOVA with the same predictors as were used for the full sample. In this less variable group, performance on the “same-different” task was no longer a significant predictor of grammar-learning ($F(1, 54) = 0.53, p = 0.47$). The Block × Context interaction was again nonsignificant ($F(12, 165) = 0.90, p = 0.55$).
The linear Block term was significant \( F(1, 165) = 73.51, p < 10^{-14} \), with performance improving overall across blocks, but the quadratic term was nonsignificant \( F(1, 165) = 1.64, p = 0.11 \), as was the cubic term \( F(1, 165) = 1.64, p = 0.20 \).

The contrast between the Smooth group and the other groups was significant \( F(1, 54) = 19.89, p < 0.0001 \), with lower performance in the former, but the Repetition Rate contrast was nonsignificant \( F(1, 54) = 2.60, p = 0.11 \). The Smooth Variance contrast was nonsignificant as well \( F(1, 54) = 0.91, p = 0.34 \). Means and standard errors are displayed in Figure 3.5b.

### 3.3.2 Discrimination Scores

#### 3.3.2.1 Full Sample

The same analyses were repeated for the confidence weighted scores described in the Procedures section for the Grammar-Learning phase (3.2.2.3). Performance on the same-different task was a
significant positive predictor of grammar-learning performance \(F(1, 114) = 14.86, p < 0.005\). The Context × Block interaction was not significant \(F(1, 345) = 0.55, p = 0.88\). The linear Block term was significant \(F(1, 345) = 85.65, p < 10^{-15}\), with an increase in performance across blocks, as was the quadratic term \(F(1, 345) = 4.21, p < 0.05\), again with performance plateauing. The cubic term was nonsignificant \(F(1, 345) = 1.92, p = 0.17\).

Once again, critically, the Smoothness contrast was significant \(F(1, 114) = 5.59, p < 0.02\), with lower performance in the Smooth groups. The Repetition Rate contrast was not significant, \(F(1, 114) = 1.00, p = 0.32\), nor was the Smooth Variance contrast \(F(1, 114) = 1.08, p = 0.30\). Means and standard errors are displayed in Figure 3.6a.

### 3.3.2.2 Trimmed Sample

The final analysis concerned the continuous scores in the trimmed sample. As with the binary outcome, when the analysis is restricted to the intermediate performers, same-different performance is not a significant predictor of grammar-learning \(F(1, 54) = 0.92, p = 0.34\). The Block × Context contrast was nonsignificant \(F(1, 165) = 1.12, p = 0.35\). The linear Block term was significant \(F(1, 165) = 77.06, p < 10^{-14}\) with performance improving over blocks. The quadratic main effect was not significant \(F(1, 165) = 2.11, p = 0.15\), nor was the cubic term \(F(1, 165) = 1.49, p = 0.22\).

As in all other analyses, the Smooth group exhibited significantly lower performance than the other Context conditions \(F(1, 54) = 15.42, p < 0.0005\). Repetition Rate was again nonsignificant, \(F(1, 54) = 1.98, p = 0.17\). The Smooth Variance contrast was nonsignificant \(F(1, 54) = 0.72, p = 0.40\). Means and standard errors are displayed in Figure 3.6b.

### 3.4 Discussion

The primary prediction in this experiment was that participants in the Smooth environment would discount the evidence for the repetition pattern, as a high rate of repetition is produced by Smoothness alone. As a result they were expected to exhibit decreased grammar-learning performance.
This prediction was supported. This effect cannot be due merely to desensitization to repetition, as greater repetition did not significantly impact learning with or without the Smoothness constraint, and in fact, the numerical difference favored greater repetition. Though it is unclear whether a significant positive effect of repetition rate might appear in a larger or less variable sample, what is clear is that the strongest predictor of performance is the qualitative shape of the context distribution.

These findings suggest that learners in this experiment are mentally modeling the alien environment, and forming hypotheses about the process that generates their input. Moreover, they appear to use this model to guide subsequent learning in the environment by assessing the evidentiary value of a cue to new potential underlying structure. Although ultimately the value of explanation may be connected to the future ability to make predictions, the absence of explicit behavioral demands frees learners to pursue a general goal of understanding the underlying nature of the environment. Here, in the Smooth environment, repetitions do not appear to be an essential component of the environment at all, whereas without Smoothness it is necessary to represent them in order to understand the distribution of intervals. Greater improvement during transfer-vocabulary blocks suggests that learners may be entertaining multiple grammar hypotheses, with the vocabulary change serving as a hint that the relevant rule is vocabulary-independent.

At first glance, the presence of differences in participants’ use of information resembles situations in which learners come to focus on features that are predictive of a relevant task outcome, filtering out redundant information. For example, Haider and Frensch (1996) trained participants to identify letter-number sequences of the form D [4] I J K and E [5] K L based on whether the bracketed numeral corresponded to the number of missing letters in the alphabetic sequence. Errors only occurred in the first two letters in the sequence. Haider and Frensch found that, while participants were initially faster to respond for shorter strings, this effect disappeared with practice, suggesting that participants were able to filter out the irrelevant letters after the first two positions. Moreover, in a test block where errors began to occur in the latter positions, errors increased, and the length
effect reappeared. Doane, Sohn, and Schriebel (1999) found similar filtering effects for judgments about novel polygons, showing additionally that the degree of filtering on a set of transfer trials increased with the difficulty of the initial task. Although many of the same mechanisms may be involved in these experiments, the nature of the learning is somewhat different. Whereas in the preceding experiments participants were engaged in a specific task all along, in the present experiment the key manipulation occurs before participants become aware of what it is they will be asked to learn. As such, it is not simply a matter of repetition being predictive of a particular response (or even of other stimulus features); rather this result suggests that learners in this experiment are creating an explanatory model of the alien environment, and forming hypotheses about how their input is being generated.

It remains to be seen whether infants at the ages studied in Dawson and Gerken (2009) will be affected by short-term exposure to a Smooth environment in the same way as the adults in this experiment. In Experiment 2, the present design is adapted to 4- and 7.5-month-old infants. Discussion of that experiment is deferred to Chapter 5, however. In Chapter 4 I discuss a formal computational model that provides an “ideal observer” gauge against which to compare the results of Experiment 1.
A BAYESIAN LEARNING MODEL FOR EXPERIMENT 1

In order to precisely quantify predicted behavior under the assumption that the smoothness constraint “competes” with repetition rules for evidence, an idealized generative statistical learning model was constructed, which “observed” an abstraction of the same melodies that the human participants heard. The use of such a model is not intended as a claim about particular psychological mechanisms involved in grammar-learning, but rather is an attempt to make explicit the inferences that result from the rational use of melodic data, given a particular set of prior beliefs about how melodies are generated. If human grammar-learning performance is affected by context in a qualitatively similar manner to the model, it will constitute evidence that human learners are influenced by some of the same factors accounted for in the model.

4.1 Model Definition

A generative probabilistic model has two components: a probability distribution (likelihood) over possible data points given a set of unknown parameters, and a prior belief distribution over possible values of those parameters. For simplicity, the likelihood component of our model entertains only the two qualitative processes used to generate melodies in the experiment. In the first process, repetition has a particular probability $p$, and the other $V - 1$ tones are equiprobable (where a uniform distribution is the case with $p = 1/V$). In the second, pitches are normally distributed around the value of the preceding tone. The model contains four free parameters, $p_1$ to $p_4$, determining the probability of repetition at each of the four sequential positions in the absence of Smoothness. A fifth parameter $h$ governs the variance of the normal distribution produced by the Smoothness
Figure 4.1: Model Likelihood Function. Probability density shown is for a tone at position $j$ after tone 2 occurs at position $j - 1$, with repetition probability $0.23 (= p_{j-1})$, variance $6 (= 1/h)$, and smooth mixture weight $0.5 (= \pi_s)$

A final parameter $\pi_s$ governs the extent to which the distribution is determined by the Smoothness constraint\(^1\). Finally, melodies are assumed to be either “grammatical” or “ungrammatical”. The two types are allowed to have different repetition probabilities at each position, but are assumed to be subject to the same Smoothness constraint, with the same mixing weight.

We use conjugate priors for all parameters, which influence inference as though additional data (melodies) had been observed prior to the experiment. The interval distribution in these “prior melodies” is a mixture of a uniform distribution and a “smooth” distribution with variance $6$, approximately equal to the empirical variance of the interval distribution in children’s folk music (Dawson, 2007). Two free parameters were varied across simulation runs. The first, $N$, determines

\(^1\)Here, $h$ is actually the “precision” of the distribution, which is the inverse of the variance. As a result, higher values of $h$ reflect a stronger Smoothness constraint.

\(^2\)note that the force of the constraint is distinct from the amount of smoothness it prescribes. The latter is governed by $h$, discussed just above.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Learned or set?</th>
</tr>
</thead>
<tbody>
<tr>
<td>${p_{1,j}}<em>{j=1}^4$ and ${p</em>{0,j}}_{j=1}^4$</td>
<td>“Pure” repetition probabilities in the absence of smoothness. Separate values for grammatical and ungrammatical items, and for each sequential position.</td>
<td>Learned</td>
</tr>
<tr>
<td>$h$</td>
<td>Precision of smoothness constraint (inverse of the variance of the interval distribution). Higher values reflect a greater tendency for small intervals.</td>
<td>Learned</td>
</tr>
<tr>
<td>$\pi_s$</td>
<td>Mixing weight of smooth process. Zero represents no smoothness constraint; one represents a distribution completely determined by smoothness.</td>
<td>Learned</td>
</tr>
<tr>
<td>$N$</td>
<td>Prior “equivalent melodies”. Higher values reduce the influence of the data relative to the prior.</td>
<td>Set</td>
</tr>
<tr>
<td>$S$</td>
<td>Prior estimate for $\pi_s$</td>
<td>Set</td>
</tr>
</tbody>
</table>

Table 4.1: Descriptions of the Model Parameters

the number of “equivalent” prior melodies encountered. The second, $S$, determines the expected weight of the Smoothness constraint.

For details on the prior and likelihood functions, see the Appendix. Example likelihood functions are shown in Fig. 4.1, and a summary of the model parameters is given in Table 4.1.

### 4.2 Simulations and Evaluation

The data “observed” by the model consisted of 200 context melodies, abstractly identical to one of the five sets encountered by the human participants, as well as the set of grammatical training melodies observed by the humans. The model’s grammatical discrimination performance was evaluated after 30, 60, 90 and 120 training melodies, corresponding to the four test blocks in the experiment. Context melodies were assumed to have a 0.5 probability of being grammatical, while the training melodies were known to be grammatical. This reflected the instructions to the human
learners, who were told that the context phase comprised both Qixian and Spy sentences, and the 50/50 distribution suggested by the distribution of alien shirts.

The bulk of the simulation process consisted of estimating a joint posterior distribution over the ten data-generation parameters, \( \{p_{1,j}\}_{j=1}^{4}, \{p_{0,j}\}_{j=1}^{4}, h \) and \( \pi_s \). This estimation was accomplished using Markov Chain Monte Carlo sampling (Gilks, Richardson, & Spiegelhalter, 1996) (see Section 4.4 for details). Simulations were run for each of the five Context conditions, with each of the four levels of training, and at three values each of the prior parameters \( N \) and \( S \). \( N \) was set to 1, 20 or 200, and \( S \) was set to 0.1, 0.5 or 0.9, where higher values reflected greater prior weight for the Smoothness constraint.

Of central interest was not what parameter values the model would infer, but how accurately it could infer the grammaticality of novel melodies. Since the sampling procedure produces a distribution of thousands of values, we can assign grammaticality probabilities to each test sentence for each set of parameter values in the sample. For each of the 24 test melodies, at each set of parameter values, the model can make a probabilistic binary decision. The mean proportions of correct responses for each simulation run are plotted in Fig. 4.2.

### 4.3 Discussion

For all parameters, the model performs more poorly in both Smooth conditions than in any of the other conditions, like the human participants. When too high a mixture weight is assigned to the Smoothness constraint (\( S = 0.9 \)), the data conveys little information about repetition probabilities, since most of the information in the interval distribution is assumed to reflect Smoothness. This makes it difficult to learn the rule in any condition, except when the prior is very weak (\( N = 1 \)). The smaller the value of \( S \), the better the model performs across conditions, as a greater proportion of the interval evidence is taken to reflect particular repetition patterns. Performance degrades with very large values of \( N \), as an overly strong prior makes the model unresponsive to the data, sticking too strongly to its “preconceptions”.
Figure 4.2: Performance of the Model by Block, Condition and Parameter Settings
The model exhibits a positive effect of repetition rate in the non-Smooth groups, with fairly large differences between the two Repetition groups. The ranks of these three conditions was the same for the humans, though the differences were not statistically significant. The strong pattern in the model may be due to its treatment of grammaticality in the Context items. The model uses approximately half of the context melodies in estimating a repetition probability for grammatical sentences. The higher the proportion of repetition there, the more valuable this signal is. The difference in humans may be smaller or nonexistent if they are using the context phase only to establish generalities about the environment, and do not retroactively treat those melodies as specifically informative about the grammar.

For the two Smooth conditions, the model predicted slightly poorer performance in the Low Variance case. The humans exhibited the opposite pattern until the last block, though again the difference was not significant. Even if human learners are sensitive to the variance of the Smooth distribution, the very small effect exhibited by the model suggests that a large sample may be needed to detect this difference.

The present model is useful for its quantitative realization of the idea that a general environmental feature can explain away evidence for a particular underlying structure in the input. The model captures a difference in human performance between two context conditions (Repetition and Smooth) that would be difficult to explain without supposing that human learners generatively model their input. If they were merely trying to learn dependencies between features and categories, the frequency of repetition alone, and not its relationship to the broader environmental structure, should be dominant in driving performance.

Although the present model captures the qualitative performance of the human participants, it should not be taken literally as an account of psychological mechanisms. The hypothesis space with which the model is endowed is extremely narrow, and is suited particularly to this task. While differences between conditions are meaningful, this model could not learn a range of other grammatical regularities that humans could likely learn. For instance, the model cannot represent
pitch contour, special roles for particular pitches, or dependencies between non-consecutive tones, all of which are potentially meaningful features of a melodic environment; nor does it attribute special salience to repetition at edges as humans seem to do (Endress, Scholl, & Mehler, 2005). Moreover, the mathematical forms of the likelihood and prior distributions were chosen largely for computational convenience. Aside from the qualitative shapes of these distributions (i.e. unimodal and smooth), their particular choices should not be taken as psychological hypotheses.

In addition, our model examines the full data set as a batch, weighing each data point equally, and representing each with perfect fidelity. This is clearly unrealistic for memory-limited humans who encounter their input over time, forming and testing hypotheses at every step. As a result, humans exhibit order-of-presentation effects that our model does not. Several recent papers have attempted to capture order effects in a Bayesian framework using sequential sampling methods (see, e.g., Doucet, Andrieu, & Godsill, 2000; Doucet, Frietas, & Gordon, 2001; Kruschke, 2006), which are a promising way forward in the effort to describe a wider variety of cognitive phenomena under the unifying computational framework of Bayesian inference.

4.4 Modeling Details

4.4.1 Model Definition

Each melody $i$ is represented as a sequence of five integers, $t_{i,1}$ through $t_{i,5}$, ranging from 1 to $V$ in ascending order of pitch, where $V$ is the number of tones in the vocabulary. For mathematical convenience, $t_{i,j}$ is determined by a probability distribution $f(t^*_{i,j})$ over the interval $[0, V]$, and a deterministic function rounding $t^*_{i,j}$ up to the nearest integer. When $j = 1$, $p(t^*_{i,1})$ is always uniform. For $j > 1$, $f(t^*_{i,j}|t_{i,j-1})$ is given by:

$$f(t^*_{i,j}|t_{i,j-1}) = \pi_s q_s(t^*_{i,j}|t_{i,j-1}) + (1 - \pi_s)q_r(t^*_{i,j}|t_{i,j-1}) \quad (4.1)$$

The functions $q_r(t^*_{i,j}|t_{i,j-1})$ and $q_s(t^*_{i,j}|t_{i,j-1})$, are the “repetition” and “smooth” densities over
tones, given by:

\[ q_r(t^*_i,j|t_{i,j-1}) \propto 1_{[t_{i,j-1}-1, t_{i,j-1}]}(t^*_i,j)p_{g,j-1} + (1 - 1_{[t_{i,j-1}-1, t_{i,j-1}]}(t^*_i,j)) \frac{1 - p_{g,j-1}}{V - 1} \]  \hspace{1cm} (4.2)

\[ q_s(t^*_i,j|t_{i,j-1}) \propto 1_{[0, V]}(t^*_i,j)h^{1/2}\exp\left(-\frac{1}{2}h(t^*_i,j - t_{i,j-1})^2\right) \]  \hspace{1cm} (4.3)

where \( 1_{[a,b]}(x) \) is the indicator function which is 1 when \( x \in [a, b] \) and 0 otherwise.

The prior distributions of \( \pi_s, h \) and \( \{p_{g,j}\} \) are given by:

\[ f_{\pi_s}(\pi_s) \propto \pi_s^{4NS}(1 - \pi_s)^{4N(1 - S)} \]  \hspace{1cm} (4.4)

\[ f_p(p_{g,j}) \propto p_{g,j}^{0.2N}(1 - p_{g,j})^{0.8N} \]  \hspace{1cm} (4.5)

\[ f_h(h) \propto h^{4N} - 1 \exp\left(-\frac{6 \times 4N}{2}h\right) \]  \hspace{1cm} (4.6)

Here, \( N \) represents the number of effective “prior” melodies, and \( S \) represents the prior strength of the smoothness constraint. The \( \pi_s \) and \( h \) parameters apply to 4 intervals per melody, for a total of \( 4N \), whereas each \( p_{g,j} \) applies to only one interval per melody, for a total of \( N \). The 6 in the \( f_h(h) \) expression represents the prior variance in the smooth distribution, fixed by the empirical distribution in Dawson (2007).

### 4.4.2 Sampling Procedure

In Gibbs sampling, the full parameter set is partitioned, and at each step a sample is taken from the conditional distribution of one block given all the others. A sample is taken for each block at each iteration, conditioning on the most recent values for the other blocks. We used separate blocks for (1) \( \pi_s \), (2) \( h \) and (3) \( \{p_{g,j}\} \). In a final block, unlabeled melodies were assigned a grammaticality label at each step and individual intervals were hard-assigned to either the repetition or smooth distributions, both according to their conditional posterior probabilities. The hard-assignment is performed to simplify sampling; each interval should be thought of as coming from a weighted
mixture of the two distributions.

Due to the truncation of the smooth distribution, the conditional posterior for $h$ is nonstandard (it would be Gamma otherwise). Therefore, a Metropolis-Hastings step (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970) was incorporated, using a Gamma as the proposal distribution.

The sampler was run for 50,000 iterations at each parameter combination. The first 25,000 iterations were discarded as “burn-in”. The remaining 25,000 samples were used to assign grammaticality probabilities to the test melodies.
5.1 Introduction

The preceding experiment was motivated by the hypothesis that cross-domain asymmetries in the ability of 7-month-old infants to learn abstract repetition rules (Marcus et al., 1999, 2007) could be attributable not to an innate bias for treating language as rule-like, but rather to a sort of higher-order learning taking place sometime before 7-months. Specifically, it was proposed in Dawson and Gerken (2009) that by 7-months of age, infants might have learned something about music that leads them to discount the occurrence of repeated tones as evidence for an underlying rule. In that paper, it was found that, as predicted, younger infants were able to learn a repetition rule instantiated in tone sequences while older infants failed, a finding taken as support for the presence of some higher-order learning in the musical domain.

The proposed content of that learning was that melodies tend to consist of small-magnitude intervals between successive pitches, and as a result, a high rate of repeated tones (being intervals of the smallest possible magnitude) is to be expected as part of the “ambient environment”, establishing a higher evidentiary bar for learning a specific repetition rule. While an analysis of the interval distributions in popular children’s melodies (Dawson, 2007) indeed confirms that smaller magnitude intervals are more frequent, establishing a correlative basis for the explaining away claim, correlation does not causation make, as the mantra goes.

Additional indirect evidence for the explaining away phenomenon proposed in Dawson and Gerken (2009) comes from Experiment 1. There, the presence of a Smoothness constraint on melodies - that is, a tendency for intervals to be small in magnitude - was shown to interfere with
adults’ ability to learn a repetition rule. This shows that, at least, the causal relationship that was left unestablished in Dawson and Gerken (2009) and Dawson (2007) shows up, at least for adult learners, and at least in a controlled setting.

To this point, the proposed explaining away account has received prima facie plausibility from the corpus analysis of Dawson (2007), the correlated discrepancy in learning observed in Dawson and Gerken (2009), and the causal link observed in a population of somewhat more advanced age in Experiment 1 here. Naturally, the next logical step is to adapt to infants the paradigm used in Experiment 1, in an attempt to observe in that population the effect of Smooth melodies seen in college students.

As is always the case when one wishes to conduct parallel experiments with adults and infants, one needs to employ slightly different methods. Naturally, the forced-choice grammaticality judgment must give way to a measure of perceived novelty such as looking time, and the entire experience must be compressed in time. An additional difficulty arises from the general impossibility of informing infants when the grammar-learning is beginning and ending. As a proxy for these explicit instructions, a change in venue between phases of the experiment is used to signal that something is changing. In addition, the five-note AABCD vs. DCBAA distinction used for the adults is replaced by simpler AAB vs. ABB grammars, paralleling those used by Marcus et al. (2007).

The preceding adaptations aside, however, the present experiment is designed to be substantively as similar as possible to the procedure in Experiment 1. In a Context phase, infants are exposed to melodies probabilistically generated according to either a Uniform, Smooth or Repetition distribution. Following this, they are familiarized a set of “grammatical” melodies adhering to either an AAB or ABB pattern. Finally, they hear a set of test melodies, half with an AAB structure, half with an ABB structure, and all with tones not heard in familiarization.

As in many other infant learning experiments, the measure of successful generalization is a difference in mean looking time between the grammatical and ungrammatical test items. The
central prediction is that learning will be more successful after exposure to either the Uniform or Repetition context than after exposure to the Smooth context. Although it is always difficult to make predictions about direction of preference for either grammatical or ungrammatical items, we follow the conventional wisdom which holds that, all else equal, a novelty preference reflects an easier learning task than does a familiarity preference. Moreover, since the 4-month-olds in Dawson and Gerken (2009) exhibited a novelty preference without having been exposed to any immediately preceding context, it would be somewhat surprising (though not impossible) if infants in the “easy” conditions here did not also exhibit a novelty preference.

In this experiment, we test the same age ranges that were studied in Dawson and Gerken (2009), namely 4-month-olds and 7.5-month-olds. The results of the experiments from that paper can be used to refine our predictions here. Since the 4-month-olds in that experiment exhibited a preference for the novel grammar, we should expect that 4-month-olds will continue to prefer the novel grammar after exposure to the Uniform and Repetition contexts. Exposure to the Smooth context, on the other hand, should result in a more difficult learning task, and hence either a familiarity preference or no preference. In either case, the critical distinction will be greater novelty preferences in the Uniform and Repetition groups than in the Smooth group.

The 7.5-month-olds in Dawson and Gerken (2009) showed no evidence of learning the repetition grammar. Taking that experiment as a baseline, exposure to the Smooth distribution should have no effect, since the pattern is already apparently too hard to learn. Moreover, the distribution of intervals in nature is Smooth anyway, and so that context should be nothing new. It is possible that, on the other hand, exposure to the Repetition or Uniform context will relieve the 7.5-month-olds of their expectation of Smoothness, at least within the local laboratory environment, which could in some sense “free” them to learn the repetition rule. If so, they should exhibit some discrimination between grammatical and ungrammatical items. Since, however, their point of reference is no preference at all, facilitation here could be reasonably expected to result in a familiarity preference, unless the effect is quite strong.
5.2 Methods

5.2.1 Participants

Thirty-nine infants between the ages of 3.5 and 4.5 months and forty-four infants between the ages of 7 and 8 months, all from the Tucson community, participated in the study. Each participant received an age appropriate book, as well as a certificate designating him or her as a graduate of the Tweety Language Development Lab, as a token of appreciation for their participation. All infants were at least 37 weeks to term, weighted at least 5 lbs. 8 oz. at birth, and had no history of speech or language problems in biological parents or full siblings.

5.2.2 Materials and Procedures

The experiment consists of a Context phase and a Grammar-Learning phase. The former consists of 3 minutes and 20 seconds of exposure to 100 melodies total, each consisting of five notes and generated according to one of three Context distributions (Uniform, Smooth or Repetition). The Grammar-Learning phase employs the Headturn Preference Procedure (Kemler Nelson et al., 1995). The familiarization phase consists of sixteen melodies all adhering to a particular repetition rule (AAB for half of infants; ABB for the other half). In the test phase, infants hear twelve test trials, half consisting of melodies from the familiarization grammar, half consisting of melodies from the opposite grammar, but all containing novel pitches. All melodies in all phases are generated using the MIDI Toolbox for MATLAB (Eerola & Toiviainen, 2004).

5.2.2.1 Context Phase: Materials

All melodies in the context phase consist of five notes, each drawn from a pool of twelve pitches, between the A below middle C (MIDI value 57) and the G♯ above middle-C (MIDI value 68). Pitches are selected probabilistically according to one of three generative processes, for a total of 100 melodies per condition. The first four tones in each melody play for 250 msec, and are followed
by 50 msec of silence. The last tone in each melody plays for 500 msec.

In the Uniform condition, each of the twelve pitches is equally probable at any point, and hence, the probability of a repetition is 1/12. In the 100 five-note melodies (i.e. 400 individual intervals) actually used, the empirical rate of repetition was 7.0%, or 28 individual intervals.

In the Smooth condition, each of the twelve tones is equally likely to occur in the first position in each melody. The probability of each subsequent tone is determined by a truncated, discretized normal distribution, centered at the previous tone, with a standard deviation of three steps. That is, the twelve tones in the vocabulary are arranged in ascending order of pitch and assigned integer values from 1 to 12. After a tone occurs, say the tone with the value 6, the probability of any given tone occurring in the next sequential position, say the tone with the value 8, corresponds to the probability of observing a value between 7.5 and 8.5 on a Normal distribution with a mean of 6 and a standard deviation of 3, and trimmed to exclude values outside the available tones (i.e., 0.5 on the low end and 12.5 on the high end). Over 100 melodies comprising 400 intervals, the empirical rate of repetition was 16.0%, which represents 64 intervals.

Finally, the Repetition condition was constructed in order to match the Smooth condition for the frequency of repeated tones. The generative probability of repetition was set equal to that in the Smooth case, but instead of the probabilities of other tones falling off gradually with increasing interval magnitude, each non-repeated tone was equally likely at every point. The resulting 100 melodies, comprising 400 intervals, contain 63 repetitions, which is 15.75% of all intervals.

For each of the three conditions, a single “pre-familiarization” trial was constructed where each of the 100 melodies is played end-to-end, with 300 msec pauses between them. The same trial is used for all infants.

5.2.2.2 Context Phase: Procedures

During the context phase, infants were seated on a caregiver’s lap in a small room. The caregiver was given headphones through which pop music was played from a portable CD player. They were
instructed not to speak to the infant, or otherwise direct her attention in any way. A projection screen was present directly in front of the infant.

Initially, the screen displayed an animated image of a cartoon baby’s face. This was present only to attract the infant’s attention prior to the start of the music. Once the infant was judged by the experimenter to be looking at the screen, the familiarization music began to play, and the image on the screen changed to a static, colored bullseye pattern. The music continued, and the bullseye remained on the screen, for 3 minutes and 20 seconds, regardless of the infant’s behavior.

Upon reaching the end of the context music, the caregiver and infant were given the opportunity to take a short break before beginning the next phase of the experiment.

5.2.2.3 Grammar-Learning Phase: Materials

Two grammar-learning conditions were crossed with the three context conditions, both employing three-note melodies, where each note plays for 562.5 msec, with 62.5 msec gaps between them. The first grammar required a repetition to be present in the first two notes, producing an $AAB$ pattern. The other grammar contained reversed versions of the same melodies, now with a repetition over the last two positions, producing a $BAA$ pattern\(^1\).

For each grammar, sixteen distinct familiarization melodies and four distinct test melodies were constructed. For the familiarization melodies, four distinct “$A$” tones and four distinct “$B$” tones were used. Each $A$ tone was paired with each $B$ tone for a total of sixteen distinct melodies. The $A$ tone was higher in pitch in half the melodies and lower in pitch in the other half. The $A$ tones were drawn from a pool of four pitches: $C_4$, $F_4$, $A_\sharp_4$ and $D_\sharp_5$ (MIDI values: 60, 65, 70 and 75). It was necessary to move two of the $B$ tones up an octave in some of the melodies in order to enforce the 50/50 split between rising and falling intervals, but this was done for the same combinations in each grammar, preserving the mirror image property. As such, the full set of pitches for the $B$

\(^1\)In most other studies in which abstract rule-learning of this sort is studied, repetition-final melodies are referred to as $ABB$. The two formulations are, of course, abstractly identical, but I depart from the usual terminology here to emphasize the fact that each melody in one grammar has a mirror-image counterpart in the other grammar.
tones was B3, C♯4, E4, F♯4, B4 and E5 (MIDI values: 59, 61, 66, 71 and 76).

Four additional test melodies were constructed for each of the two grammars, in a similar way. Two new A tones and two new B tones were used, with each of the four combinations represented. As in the familiarization melodies, the A tone was higher than the B tone in half the melodies, and lower in the other half. In addition to individual pitches being completely new from familiarization to test, no intervals were shared with any of the familiarization melodies. The A tones consisted of A3 and D4 (MIDI values: 57 and 62), and the B tones consisted of G♯3 and G4 (MIDI values: 56 and 67).

A single familiarization trial was constructed for each grammar, containing three blocks, where each block contains all sixteen melodies once in a different random order. The same random orders were used for the two grammars, so that the first melody in the BAA sequence was the mirror image of the first melody in the AAB sequence, and so on. Gaps of 625 msec were inserted between each melody, with no distinct marker separating blocks. In total, the familiarization sequences spanned exactly 2 minutes.

Two test trials were constructed for each grammar, according to the same procedure as the familiarization trials: each of the four test melodies occurred three times, randomized within blocks, with matched random orders between grammars. The only difference between the two test trials of a given grammar was the order of the melodies.

5.2.2.4 Grammar-Learning Phase: Procedures

Familiarization and test subphases proceeded according to the Headturn Preference Procedure (Kemler Nelson et al., 1995). As in the context phase, caregivers listened to music over headphones and were instructed not to direct the infant’s attention in any way. During the familiarization phase, a light flashed in front of the infant until the experimenter judged the infant to be looking at it. After the infant was judged to be looking at the center light, it turned off and a blinking light came on on either the left or right side at random. The side light remained on until the infant first fixated
on it and then looked away for 2 contiguous seconds, at which point it was extinguished and the center light resumed blinking. This cycle continued throughout the familiarization phase. While the behavior of the lights was contingent on the infant’s looking behavior, the music continued to play uninterrupted, regardless of where the infant was looking.

Immediately after the familiarization phase concluded, the test phase began. The behavior of the lights was contingent on the infant’s looks just as in the familiarization phase, but now the sound was contingent as well. A test trial began to play each time the infant looked at the blinking side light, and stopped after a contiguous 2 second look-away or after 30 seconds had elapsed, whichever occurred first. Each infant heard three blocks of four test trials, with each of the four trials occurring once per block, in a different random order in each block and for each infant. The side on which the light blinked was chosen independently for each trial, and independently of the auditory stimulus.

5.3 Results

Although the theoretical prediction is qualitatively the same for both age groups — namely, that learning the repetition rule will be more difficult after exposure to Smooth melodies than after exposure to either the Uniform context or the Repetition context — the behavioral predictions are different for 4-month-olds and 7.5-month-olds, on the basis of baseline learning behavior exhibited in Dawson and Gerken (2009). As discussed previously, since 4-month-olds can extract repetition generalizations, exhibiting a novelty preference with no context manipulation, they should be expected to exhibit the same preference after exposure to non-Smooth melodies. If concentrated exposure to melodies subject to a Smoothness constraint interferes with learning of a repetition rule, 4-month-old infants in the Smooth group should exhibit either a familiarity preference or no preference. On the other hand, 7.5-month-olds in the Smooth group should exhibit no preference, just as they do without any context manipulation, while possibly exhibiting a preference in the other conditions.
5.3.1 Exclusion Criteria

In order to interpret looking time during a given test trial as an indication of grammar-related surprise (or familiarity), it is necessary for enough of the trial to have played to determine whether it is an \( AAB \) trial or a \( BAA \) trial. Since a single melodic sequence lasts approximately 2 seconds, an infant who looks away after less time than this cannot have received enough information to determine whether the trial contained grammatical or ungrammatical melodies. Consequently, all trials with looking times below 2 seconds were excluded from analysis. Moreover, all data from any infant who did not have looking times of at least 2 seconds for at least three of the six test trials in each grammar was excluded.
5.3.2 Planned Analyses

Since different infants had different numbers of test trials included in the analysis, the resulting design was unbalanced. Since traditional Analysis of Variance assumes equal numbers of observations in each cell, we turn to a General Linear Mixed Model to analyze the looking time data from this experiment. Specifically, the lme procedure in the nlme package in R (Pinheiro, Bates, DebRoy, Sarkar, & R Core team, 2009) was used to compute Maximum Likelihood estimates and standard errors of the model parameters.

Data from the two age groups were analyzed separately, with Context condition (Smooth, Uniform and Repetition) as a between-subjects fixed factor, and Grammaticality (Grammatical vs. Ungrammatical) and test Block (1, 2 and 3) as within-subjects fixed factors. Variation among infants’ overall looking times was captured using a random intercept term for Subjects.

The Context factor was broken down into two planned orthogonal contrasts. The primary
contrast of interest was between the Smooth group, on the one hand, and the other two groups on the other. The second contrast was between the Repetition and Uniform groups. Two Block contrasts are potentially of interest in assessing changes in looking patterns during the course of the test phase. The first of these compares looking behavior during and after the first test block. The second compares behavior before and during the last test block.

### 5.3.2.1 4-month-olds

A likelihood ratio test was used to test whether including Familiarization condition (AAB vs. ABB) as a predictor improved model fit. Overall, the model with Familiarization condition, and all higher order interactions involving Familiarization condition, included as predictors did not fit significantly better than the analogous model with all terms related to Familiarization condition omitted ($\chi^2(18) = 24.12, p = 0.15$). Therefore, the other effects were evaluated for the simpler model.

The three-way interaction between Grammaticality, Block and the Smooth contrast was not significant ($F(2, 366) = 0.70, p = 0.49$), nor was the interaction between Grammaticality, Block and the Repetition vs. Uniform contrast ($F(2, 366) = 1.06, p = 0.35$).

The critical interaction, between the Smooth Context contrast and the Grammaticality effect, was significant ($t(366) = 2.23, p < 0.03$), with greater preferences for the novel test items in the Repetition and Uniform groups than in the Smooth group. The interaction between the Repetition vs. Uniform contrast and the Grammaticality effect was significant as well ($t(366) = 2.18, p < 0.04$), with greater preferences for the ungrammatical items in the Uniform group than in the Repetition group. The only other interaction that approached significance was the interaction between Grammaticality and the Block 3 contrast. Across Context conditions, greater novelty preferences were exhibited in the third test block ($t(366) = 1.71, p = 0.09$).

Main effects are difficult to interpret for the full model in the presence of the Grammaticality × Context and Grammaticality × Block interactions. Overall looking times by Context and Block
are displayed in Fig. 5.2. Mean preferences by Context and Block are displayed in Fig. 5.1.

Due to the significant interactions between Grammaticality and Context, data for each Context group was analyzed separately, with Block and Grammaticality included as predictors. The results for each group are reported in turn.

For the Uniform group \((n = 14)\), the Grammaticality \(\times\) Block interaction was not significant \((F(2, 127) = 0.99, p = 0.38)\). The main effect of grammaticality was significant \((t(127) = 2.42, p < 0.02)\), with infants exhibiting an overall preference for the novel test items \((M_{\text{gramm.}} = 9.7s, M_{\text{ungramm.}} = 12.6s)\). The main effect of Block was also significant \((M_{B1} = 14.1s, M_{B2} = 9.8s, M_{B3} = 9.5s, F(2, 127) = 6.98, p < 0.005)\).

For the Smooth group \((n = 13)\), the Grammaticality \(\times\) Block interaction was not significant \((F(2, 125) = 0.60, p = 0.55)\). The main effect of Grammaticality was significant \((t(125) = 2.05, p < 0.05)\), but the preference was for the familiar test items \((M_{\text{gramm.}} = 11.2s, M_{\text{ungramm.}} = 8.9s)\). Overall looking times differed significantly across blocks \((M_{B1} = 11.9s, M_{B2} = 8.7s, M_{B3} = 9.4s, F(2, 125) = 3.47, p < 0.04)\).

For the Repetition group \((n = 12)\), the interaction between Grammaticality and the Block 3 contrast was significant \((t(114) = 2.13, p < 0.04)\). The overall effect of grammaticality was not significant \((M_{\text{gramm.}} = 12.5s, M_{\text{ungramm.}} = 11.2s, t(114) = 0.65, p = 0.52)\), nor was the effect of Block \((M_{B1} = 12.57, M_{B2} = 11.19, M_{B3} = 11.84, F(2, 114) = 0.53, p = 0.59)\).

Given the significant Grammaticality \(\times\) Block interaction in the Repetition group, preferences in this group were analyzed for the first two blocks separately from the last block. In the first two blocks, a marginally significant preference for the familiar items was observed \((M_{\text{gramm.}} = 13.5s, M_{\text{ungramm.}} = 10.2s, t(76) = 1.87, p = 0.07)\). In the last block, looking times were longer for novel items \((M_{\text{gramm.}} = 10.3s, M_{\text{ungramm.}} = 13.3s)\), but despite the absolute preference being numerically comparable in magnitude to that in the first two blocks, it failed to reach significance here \((t(29) = 1.20, p = 0.24)\), due to greater variability and fewer data points.
5.3.2.2 7-month-olds

The same analyses were carried out for the 7-month-olds.

A likelihood ratio test comparing models with and without Familiarization condition included as a predictor revealed that the fit was not significantly improved by its inclusion ($\chi^2(18) = 15.75$, $p = 0.61$), and so it was omitted from subsequent analyses.

The three-way interaction between Grammaticality, Block and the Smooth contrast was not significant ($F(2, 431) = 0.18$, $p = 0.83$), nor was the interaction between Grammaticality, Block and the Repetition vs. Uniform contrast ($F(2, 431) = 0.49$, $p = 0.61$). The interaction between Grammaticality and the Smooth Context contrast was not significant ($t(431) = 0.83$, $p = 0.41$), nor was the interaction between Grammaticality and the Repetition vs. Uniform contrast ($t(431) = 0.34$, $p = 0.73$). The interaction between Grammaticality and the Block 3 contrast was significant ($t(431) = 2.14$, $p < 0.04$), with greater familiarity preferences overall in Block 3. The interaction between the Smooth context contrast and the Block 1 contrast was significant ($t(431) = 2.05$, $p < 0.05$). Looking times declined after Block 1 in the Repetition and Uniform groups, but increased in the Smooth group. No other interactions or main effects approached significance. Mean preferences by Context condition and Block are shown in Fig. 5.1.

Since there was a significant interaction involving the Smooth contrast, data was analyzed for the Smooth group separately from the other two groups, with Block and Grammaticality included as predictors for the Smooth group, and Block, Grammaticality and Context (Repetition vs. Uniform) included as predictors for the other two groups.

For the Smooth group ($n = 14$), the Grammaticality × Block interaction was not significant ($F(2, 128) = 0.47$, $p = 0.62$). The overall effect of Grammaticality was not significant ($M_{gramm.} = 7.8s$, $M_{ungramm} = 7.9s$, $t(128) = 0.32$, $p = 0.75$). The difference in overall looking times across Blocks was not significant ($M_{B1} = 6.9s$, $M_{B2} = 9.2s$, $M_{B3} = 7.3s$, $F(2, 128) = 2.16$, $p = 0.12$).

For the Repetition and Uniform groups ($n = 18$ and 12, respectively), the three-way interaction
between Grammaticality, Block and Context was not significant ($F(2, 303) = 0.48, p = 0.62$). The interaction between Grammaticality and the Block 3 contrast was significant ($t(303) = 2.04, p < 0.05$), with greater familiarity preferences in the third block. Overall, the main effect of grammaticality was not significant ($t(308) = 1.06, p = 0.29$), nor was the main effect of Block ($M_{B1} = 8.2s, M_{B2} = 7.6s, M_{B3} = 7.4s, F(2, 303) = 0.79, p = 0.32$).

Due to the marginally significant Grammaticality $\times$ Block interaction for the non-Smooth groups, looking times were analyzed for the first two blocks separately from the last block, with Grammaticality and Context (Repetition vs. Uniform) as predictors. In Blocks 1 and 2, the interaction between Grammaticality and Context was not significant ($t(199) = 0.16, p = 0.87$), nor were the main effects of Grammaticality ($M_{gramm.} = 8.0s, M_{ungramm.} = 7.9s, t(199) = 0.16, p = 0.87$) or Context ($M_{Rep.} = 7.9s, M_{Unif.} = 7.9s, t(30) = 0.10, p = 0.92$).

In Block 3, the interaction between Grammaticality and Context was also nonsignificant ($t(78) = 0.75, p = 0.46$), as was the main effect of Context ($M_{Rep.} = 7.1s, M_{Unif.} = 7.7s, t(30) = 0.56, p = 0.58$). Now, however, a significant preference for grammatical items was exhibited ($M_{gramm.} = 8.1s, M_{ungramm.} = 6.5s, t(78) = 2.22, p < 0.03$).

### 5.3.3 Interim Discussion and Post Hoc Analyses

The primary prediction was supported for the 4-month-olds: namely, infants in the Repetition and Uniform conditions displayed greater preferences for the ungrammatical items than their counterparts in the Smooth condition. However, when the Uniform and Repetition Context groups are analyzed individually, only the Uniform group shows a significant overall preference, and indeed, when Block is ignored, the Repetition group appears more similar to the Smooth group. Collapsing across Blocks, no differences whatsoever were seen in the 7-month-olds. However, some unexpected effects of Block appeared in both ages, with the last block of test trials appearing to stand apart from the others, yielding preference patterns that are closer to predictions than those observed in the early test trials. This is suggestive of continued learning during the test phase, and may
even reflect infants entertaining multiple hypotheses initially, and using the test data to distinguish between them. As a matter of thoroughness, a set of post hoc statistical analyses were conducted, to explore these unexpected findings more deeply. Since the hypotheses informing the analyses that follow were arrived at after the data was examined, the results and interpretations in this section should be taken with a grain of salt: at most, they raise some interesting questions for future research.

5.3.3.1 4-month-olds

During the early test blocks, the Repetition group appeared similar to the Smooth group, displaying a preference for the grammatical test items, whereas the Uniform group preferred the ungrammatical items. During the last block, however, the Repetition group reversed its mean preference, clustering with the Uniform group in preferring the ungrammatical items, while the Smooth group continued to show a familiarity preference. This reversal may reflect the ambiguity inherent in the Repetition context: it shares features with both the Smooth and the Uniform contexts, and its high rate of repeated tones could potentially be interpreted either as a part of the “background noise”, as in the Smooth case, or as a signal that repetition is a meaningful feature of the environment. It is possible that, upon hearing the grammatical and ungrammatical sequences side by side at test, this ambiguity is resolved in favor of the latter explanation, and by the last test block, infants in the Repetition group shift their focus to the ungrammatical items.

One piece of evidence supporting this possibility is the consistently high looking times across blocks in the Repetition condition: Where the infants in the Smooth condition exhibit significantly decreased looking times after the first block, infants in the Repetition condition are still looking as long in the last block, on average, as the Smooth infants were in the first (within 1/20 of a second). In the Smooth context, on the other hand, the hypothesis that repetition is meaningful is present, but has less evidence in its favor prior to exposure to the grammar; hence the more uncertain familiarity preference persists throughout the test trials.
Looking times in the last block \( (n = 127) \) were analyzed with Grammaticality and Context as predictors, using the same Context contrasts as in the preceding analyses. As in the overall analysis, the interaction between Grammaticality and the Smooth contrast was significant \( (t(93) = 2.26, p < 0.03) \), with greater novelty preferences in the two non-Smooth groups. Now, however, the interaction between Grammaticality and the Repetition vs. Uniform contrast is nonsignificant \( (t(93) = 0.41, p = 0.69) \). This pattern of results is similar to that of the adults in Experiment 1, with adult performance translated to infant novelty preferences.

When Block 3 looking times for the Repetition and Uniform groups are analyzed with Grammaticality and Context as predictors, the interaction is not significant \( (t(59) = 0.42, p = 0.68) \), nor is the main effect of Context \( (t(24) = 1.27, p = 0.21) \); however, there is a significant main effect of Grammaticality \( (t(59) = 2.32, p < 0.03) \), with longer looking times for novel test items \( (M_{gramm.} = 8.8s, M_{ungramm.} = 12.2s) \).

### 5.3.3.2 7-month-olds

As discussed in the introduction to the present chapter, given that 7-month-olds do not appear to learn repetition rules in the absence of context exposure (Dawson & Gerken, 2009), it is reasonable to expect a greater level of difficulty overall in the present experiment. In contrast to the 4-month-olds, who exhibited a novelty preference in the 2009 experiments, and who were expected to again exhibit a novelty preference in the ‘easy’ conditions here, it was suggested in the introduction that facilitation for the 7-month-olds could well produce a familiarity preference.

As with the 4-month-olds, the 7-month-olds appear to behave somewhat differently in the last test block than in the first two, with results there appearing closer to the original predictions. As was suggested for the 4-month-olds, the emergence of a significant preference in the last block for the two non-Smooth groups may reflect learning during the test phase. No such preference appeared for the Smooth group. In the last block, the pattern of results is similar to what was predicted: Although the interaction between Smoothness and Grammaticality is not significant,
the two non-Smooth groups appear to exhibit a preference, whereas the Smooth group does not.

5.3.3.3 Comparison Across Ages

It is somewhat difficult to look for similarities in the context effect between the two ages, since 7-month-olds have a more difficult time learning the repetition rule to begin with. Although both age groups exhibited behavior consistent with predictions in the last test block, the reversal in their direction of preference makes simply pooling the data fairly uninformative. One approach that may make some sense, however, is to reverse the coding of the test items for one of the age groups, so that novelty preferences for the 4-month-olds are treated the same as familiarity preferences in the 7-month-olds, as far as the statistical model is concerned. This is similar to using test item novelty as a predictor variable in place of the particular test grammar: For half of the infants, $AAB$ trials are novel; for the other half they are familiar.
Grammaticality was recoded as described above, and looking times for both ages were analyzed in the same manner as in previous analyses, with Grammaticality, Context, Block, and Age as predictors. The same contrasts were used for Context and Block as before.

The four-way interaction was not significant ($F(4, 805) = 0.95, p = 0.44$). The interaction between Grammaticality, Age and the Smoothness contrast is nonsignificant ($t(805) = 1.31, p = 0.19$), as is the interaction between Grammaticality, Age and the Repetition vs. Uniform contrast ($t(805) = 1.59, p = 0.11$) suggesting that Context has a similar effect on preferences for both ages, albeit in opposite directions. No other three-way interactions were significant either.

A significant two-way interaction between Grammaticality and the Smoothness contrast ($t(805) = 2.31, p < 0.03$) reflects weaker overall preferences in the Smooth condition. The interaction between Grammaticality and the Repetition vs. Uniform contrast was also significant ($t(805) = 2.01, p < 0.05$) reflecting stronger preferences in the Uniform condition, possibly due to ambiguity in the Repetition distribution, as suggested above. The interaction between Grammaticality and the Block 3 contrast was marginally significant ($t(805) = 1.84, p < 0.07$), with stronger preferences in Block 3 overall.

The main effect of the Smoothness contrast on overall looking times was marginally significant ($t(805) = 1.69, p < 0.10$), with shorter looking times in the Smooth condition. Overall looking times did not significantly differ between the Repetition and Uniform conditions ($t(805) = 0.68, p = 0.50$). The interaction between Age and the Smoothness contrast was not significant ($t(805) = 0.38, p = 0.70$), nor was the interaction between Age and the Repetition vs. Uniform contrast ($t(805) = 0.43, p = 0.67$). This pattern suggests that the two non-Smooth conditions induce a greater level of engagement than does the Smooth condition, perhaps a measure of infants’ ‘metacognitive’ ability to monitor how much they are learning.

When Block 3 is isolated, and looking times are analyzed using Context, Age and the recoded Grammaticality variable, the picture is quite consistent with predictions. The only significant interaction is the two-way interaction between Grammaticality and the Smoothness contrast.
\[(t(211) = 2.38, \ p < 0.02),\] with stronger preferences in the non-Smooth conditions. In particular, the interaction between Grammaticality and the Repetition vs. Uniform contrast is not significant \[(t(211) = 0.69, \ p = 0.49).\] The main effect of Age is significant \[(t(211) = 3.37, \ p < 0.001),\] with shorter overall looking times by the 7-month-olds, which is not surprising. However, Age does not enter into any other significant interactions (all \(ps > 0.2\)). In fact, when Age is treated as an ‘adjustment’ variable, included only as a main effect, the model fit does not significantly suffer (Likelihood Ratio \(\chi^2(5) = 3.76, \ p = 0.58),\) suggesting that the effect of Context, both on overall looking times and grammaticality preferences, is indistinguishable for the two ages, up to the reversal in preference direction.

Overall preferences by Age, collapsed across blocks and for Block 3 alone, are displayed in Figure 5.3.

5.4 Discussion

The infant data supports the explaining away account, particularly in the last block of test trials. Four-month-olds who are familiarized with context melodies in which the distribution of intervals is uniform are able to learn the AAB/ABB grammar, exhibiting a looking time preference for the ungrammatical test items, just like infants of the same age who do not receive any context manipulation in the laboratory (Dawson & Gerken, 2009). Those familiarized with context melodies that are subject to a Smoothness constraint, while still able to learn the AAB/ABB rule, exhibit a reversal in their preference, likely reflecting increased difficulty and/or uncertainty associated with extracting the repetition rule.

Contrary to the initial prediction, infants in the Repetition condition behave more similarly to those in the Smooth condition early on in the test phase; though by the last block of test trials they are exhibiting a novelty preference along with the Uniform group. This reversal of preference

\(^2\)As an ancillary point, the 4-month-old data shows that the rule-learning observed in Dawson and Gerken (2009), which used an AAB vs. ABA contrast, applies equally well to the potentially more subtle AAB/ABB contrast. While AAB and ABA can be distinguished merely on the basis of presence or absence of an immediate repetition, distinguishing AAB from ABB requires attention to the sequential position of that repetition.
requires additional study before firm conclusions can be made, but if the effect is a stable one, it would seem to suggest two things: First, greater ambiguity is present in the Repetition condition than in the Uniform condition, perhaps simply due to the noise introduced by extra repetitions, or possibly due to its similarity to the Smooth condition. In addition, this ambiguity is resolved more quickly, relative to the Smooth condition, with the side-by-side exposure to the two grammars that occurs in the test phase.

It is interesting to note that 4-month-olds in the Uniform condition appeared most similar to 4-month-olds in Dawson and Gerken (2009), who engaged in grammar-learning without prior context exposure. The two are not directly comparable, due to numerous differences in the experimental procedure between the two studies (infants in Dawson and Gerken heard $AAB$ or $ABA$ sequences, while infants here heard $AAB$ or $ABB$; moreover, infants in the present experiment were engaged for several minutes more than those in Dawson and Gerken prior to hearing the test items, due to the presence of the Context phase). Nonetheless, one might suppose that, absent context exposure in the laboratory, infants draw on their previous months of experience with natural music.

As observed in Dawson (2007), natural music is Smooth, provided intervals are measured with respect to the diatonic scale. Hence, at first glance, it is surprising that the 4-month-olds in Dawson and Gerken would not have exhibited a pattern of results parallel to those in the Smooth condition here, rather than the Uniform condition. Surely, if 3 minutes of exposure to a novel context in the laboratory is sufficient to establish a model, 4 months of exposure would be enough to do the same. However, as I observed in Dawson (2007), if one does not yet possess a concept of diatonic key, the distribution of intervals, collapsing across keys, cannot be fit well by a smooth distribution; in many ways the distribution is closer to the Repetition condition used here. In light of this fact, 4-month-olds should be unlikely to have a Smoothness model ready to apply, prior to entering the laboratory.

As predicted, seven-month-olds in the Smooth condition exhibited no ability to distinguish grammatical from ungrammatical items after familiarization with an $AAB/ABB$ grammar. Ten-
tative evidence from the last block of test trials suggests that exposure to the intervals in the other
two conditions is sufficient to allow seven-month-olds to “unlearn” whatever is preventing them
from learning the repetition rules. However, they exhibit a familiarity preference here, suggesting
that learning the repetition grammar is still more difficult than it is for the 4-month-olds. ‘Un-
learning’ is likely more difficult than initial learning. Again, the effect of Block was not predicted
in advance, and hence will need to be replicated (perhaps using a habituation design) before any
firm conclusions can be drawn; but it offers at least a hint that a prior expectation of Smoothness
is contributing to 7-month-olds’ failure to learn repetition rules, and that that expectation can be
unlearned with enough experience in a non-Smooth environment.

While the evidence for the explaining-away hypothesis would have been stronger had the patterns
seen in the final block of test trials appeared throughout the test phase, it is somewhat reassuring
that the same block effects appear for both age groups. This symmetry somewhat reduces the
likelihood that the effects are statistical flukes. Nonetheless, additional data is needed.

Clearly there are differences in the way infants and adult approach the grammar learning problem
explored here. Perhaps the most obvious difference is that adults are explicitly aware that they are
performing a psychology experiment. For this reason, perhaps as well as the fact that they have
somewhat more fully developed executive brain functions, they are likely to have an easier time
insulating the “alien environment” from their typical experience, and an easier time segregating
the context phase from the grammar-learning phase. This last difference in particular may help
to explain why, for example, 4-month-olds have a more difficult time learning in the Repetition
condition than in the Uniform condition. Or, as mentioned above, the lesser context exposure on
the part of the infants may limit infants’ differentiation of the Repetition and Smooth distributions.

It would be over-interpreting the data to suppose that the fact that 4-month-olds and adults
successfully learn the repetition rule, while 7.5-month-olds appear not to do so (at least not without
exposure to counterexamples in the test phase) reflects some deeply meaningful U-shaped develop-
ment curve. The apparent nonlinearity may simply be due to differences in the nature of the task.
Nonetheless, it is not unreasonable to suppose that adults can more easily “borrow structure” from other domains: Even if a repetition rule is very unnatural in music, to the point where 7.5-month-olds would have great difficulty learning it under any normal circumstances, adults may be able to make connections with other domains in a way that infants are not yet capable of doing.
CHAPTER 6

EXPERIMENT 3

6.1 Introduction

The central intuition behind Experiment 1, and the accompanying Bayesian model, is that there exist at least two ways to represent tone events, and the relationships among pitch values. In a discrete representation, a sort of arbitrariness of the sign (Saussure, 1916) might be expected to hold, with some degree of separation between surface similarity and functional similarity. In particular, tones that are nearby in frequency space need not play similar structural roles in a melody, whereas functionally similar tones (e.g., that are part of the same chord) may be several steps away. In a continuous representation, on the other hand, the underlying continuous dimension of frequency becomes important. It is this latter representation that is necessary to explain the tendency for melodies to contain small intervals between temporally successive pitches. While these two representations are undoubtedly mixed in natural music, it was shown in Experiments 1 and 2 that their relative importance could be shifted by manipulating the statistical information contained in the distribution of intervals.

The primary prediction in the first two experiments was that the structure of the musical context to which participants are exposed can influence which of these general representational schemes dominates. In that experiment, after exposure to one of three types of context distribution, participants were asked to learn about an alien dialect in which grammaticality depended on the presence of a pair of repeated tones at a particular point in the melody. There, it was predicted that representing tones as discrete types, rather than elements on a continuum, would be more conducive to learning the grammar. Indeed, when the structure of the context included a Smoothness constraint
in which intervals between consecutive pitches tended to be small in magnitude, a property that requires a representation of an underlying pitch continuum, participants had a more difficult time learning the repetition grammar than when no evidence for an underlying continuum was present.

If learners are indeed entertaining two competing fundamental representations of the tone environment, one discrete and abstract and one continuum-based, then the relative performance under the various context conditions should be a function of the nature of the grammar being learned. It was claimed that the reason participants in Experiment 1 performed more poorly in the Smooth context condition was that the context melodies led them to represent the continuum underlying the tone vocabulary, and that this masked the presence of repetitions. One might have the objection, however, that smooth melodies are simply more compelling to listen to, and that learners in that environment devoted more cognitive resources to the grammatically irrelevant context, leading to increased interference with the grammar-learning task. Or, instead, perhaps the Smooth melodies were in fact less interesting due to greater predictability, and this led to participants being fatigued by boredom when the time came to learn the grammar. In either case, the performance discrepancy would be attributed not to deep differences in the way learners represented the tone environment but simply to surface-level processing factors. If either of these processing explanations is driving the effect seen in the first experiment, then participants in the Smooth context should perform more poorly regardless of the nature of the grammar being learned. That is, compatibility of learners representation with the task should not be a factor.

If, however, the tone context is driving participants to employ different representations of the environment, then the use of a grammar that depends explicitly on the existence of an underlying pitch continuum should reverse the pattern of performance across context conditions. Namely, encountering evidence for such a continuum in the context phase, as should occur in the Smooth condition, should assist, rather than interfere, with learning. If instead the difference in Experiment 1 is due to general processing factors resulting in the different contexts being differentially engaging, a similar pattern of results should be produced when the nature of the grammar is altered. The
present experiment is designed to differentiate between these two hypotheses.

For Experiment 2, the grammaticality of training and test melodies will depend on pitch contour: the overall “shape” of the melody if it were graphed with time on the horizontal axis and pitch height on the vertical axis. For half of participants, grammatical melodies will have an “arc” shape, moving in one direction (either rising or falling in pitch) for the first two intervals, and in the opposite direction for the second half. Abstractly, these melodies can be said to have either a “V” shape, or the vertical inversion thereof. For the other half of participants, grammatical melodies have a “zig-zag” shape, with alternating rising and falling intervals. That is, if the second note is higher in pitch than the first, then the third note must come back down, and vice-versa; then the fourth note must go back up, and so on. These melodies can be imagined abstractly as having the shape of an italicized “N”, or its mirror image. As in the first experiment, melodies from the opposite grammar serve as ungrammatical foils in the test phase for both groups of participants.

Since these contour grammars can only be learned if the learner has access to pitch height relations, a learner who believes strongly that the alien language employs the property of “arbitrariness of the sign”, in which surface similarities are irrelevant to underlying functional significance, will have no way to draw the relevant generalization across grammatical melodies. As such, learners in the Repetition and Uniform context conditions, in which prior experience supports such an arbitrariness property, should have a more difficult time learning the contour grammar than learners exposed to Smooth melodies, which support the notion that the pitch continuum matters.

6.2 Methods

6.2.1 Participants

Seventy-two University of Arizona undergraduates participated in the study for course credit. An additional eleven participated but were excluded from analysis based on their failure to score above chance on a melodic-discrimination screening task.
6.2.2 Materials and Procedures

The overall structure of the experiment is identical to Experiment 1. Participants first completed a context phase, followed by four blocks of training and test items, during which they were given the task of learning the grammar of an alien dialect. As before, all “sentences” consisted of five tones, generated using the FM Synthesizer in the MIDI Toolbox for MATLAB (Eerola & Toiviainen, 2004), which produces a horn-like sound. The first four notes are 250 msec each, with 50 msec gaps after each one. The last note is 500 msec.

6.2.2.1 Procedures and Materials: Context Phase

The context phase proceeds identically to Experiment 1. Twenty-four participants were assigned to each of the Repetition, Smooth or Uniform conditions. Since the Variance manipulation had no observed effect in the first experiment, only the High Variance varieties of the Repetition and Smooth conditions were used. As before, in all cases, context melodies are drawn from a vocabulary of six tones: A3, A♯3, C♯4, E4, G4 and G♯4 (MIDI values 57, 58, 61, 64, 67 and 68).

6.2.2.2 Procedures: Grammar-Learning Phase

The procedures for the Grammar-Learning Phase were identical to Experiment 1. This stage consisted of four training blocks, each immediately followed by a test block. In each training block, participants heard thirty “grammatical” sentences in random order while an image of four star-chested aliens is displayed. After each training block, participants hear twenty-four test sentences, half grammatical. After each sentence, participants make a continuous grammaticality judgment by clicking on the response line.

6.2.2.3 Materials: Grammar-Learning Phase

The only difference between Experiment 1 and Experiment 2 is the nature of the grammar on which participants are trained and tested during the grammar-learning phase. Whereas in Experiment 1,
grammaticality depended on the presence of a repeated tone pair at either the beginning or the end of a tone sequence, in the present experiment, it is the pattern of rising and falling intervals that defines the grammar. The Qixian and spy sentences are again five tones in length. Each participant is trained using one of two six-tone vocabularies\(^1\). The first (V1) contains the tones A\(_3\), C\(_4\), C\(_\#\)\(_4\), D\(_\#\)\(_4\), F\(_\#\)\(_4\) and G\(_4\) (MIDI 57, 60, 61, 63, 66 and 67); the second (V2) contains the tones A\(#3\), B\(_3\), D\(_4\), E\(_4\), F\(_4\) and G\(_\#\)\(_4\) (MIDI 58, 59, 62, 64, 65 and 68). Each set shares three tones with the context vocabulary.

The first grammar contains melodies with an arc shape. These are melodies with either a \(++++\) interval sequence or a \(----\) sequence (where + denotes a rising interval and - denotes a falling interval). The opposite grammar contains zig-zag melodies, which have either the interval sequence \(-+-+\), or the interval sequence \(-+-+\). A schematic representation of the two grammars is

\(^1\)It was necessary to use six-tone, rather than five-tone, vocabularies in order to generate a large enough pool of sentences, since in, the grammars used here, all five tones are distinct in a given sentence.
depicted in Figure 6.1. Training sets consisted of 60 distinct melodies of each grammar/vocabulary combination, half with each specific rising-falling profile. Of these, 30 were chosen at random to be used in training blocks 1 and 2; the other 30 were used in blocks 3 and 4. Test sets consisted of 12 distinct melodies from each grammar-vocabulary combination. In odd-numbered test blocks, grammatical and ungrammatical test items from the training vocabulary were used; in even-numbered blocks, grammatical and ungrammatical test items from the opposite vocabulary were used.

6.3 Results

The primary prediction was that exposure to the Smooth context condition would reinforce the notion that tones in the alien language were meaningfully organized along a pitch continuum, whereas exposure to one of the other contexts would leave learners with the impression that any functional similarity between two pitches is unrelated to their proximity on the surface dimension of frequency. As a result, learning to distinguish melodies based on their contour should be more difficult after exposure to the Repetition and Uniform contexts.

It is unclear what prediction to make about the relative performance of the Repetition and Uniform groups. On the one hand, the presence of extra repetitions could strengthen the support for a discrete organization of pitch, at the expense of continuous properties like contour. In Experiment 1, the Repetition groups outperformed the Uniform group, consistent with this hypothesis. On the other hand, if, rather than comparing a Smoothness model with a Repetition-based model, one simply attempts to determine how well the Smoothness model fits the data, the peaked interval distribution in the Repetition condition may afford a better fit than the Uniform distribution, depending on how strong the tendency for small intervals is expected to be under the Smooth model. In this case, performance in the Repetition condition may fall between the Uniform and Smooth conditions. Given this ambiguity, we hold off on making any predictions about the relative performance in the two non-Smooth conditions.

As in Experiment 1, two sets of analyses were conducted: one on the data from the full set
of participants, and another on only the data from participants with accuracy scores between the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles for their context condition. All 18 participants who were excluded for low performance were at floor, by the definition used in Experiment 1 (namely, fewer than 57 correct responses overall). Of the 18 participants in the uppermost quartile, 5 were at ceiling. Of the 36 participants included in the trimmed analysis, 18 were at floor, and none were at ceiling. Although performance in this experiment is much less symmetrical with respect to floor and ceiling performance, in order for the comparison with Experiment 1 to be made based on roughly the same segment of the population, the quartile-based exclusion criterion was retained nonetheless.

6.3.1 Number Correct

Both the binary (correct vs. incorrect) and continuous (confidence-weighted) responses were analyzed. We report the results from the binary scores first.
6.3.1.1 Full Sample

The total number of correct responses was computed for each participant at each block and these scores were entered into an ANCOVA with between-subjects factor Context Condition (three levels: Smooth, Repetition and Uniform), within-subjects factor Block (1 through 4), and a covariate based on the number of correct same-different responses during the probe trials in the Context phase. These scores were standardized to z-scores within each Context condition. Means and standard errors for each block and each group are displayed in Figure 6.2a.

Two orthogonal planned contrasts were used for the Context factor. The first compared the Smooth group to the other two distributions; the second compared the Repetition group to the Uniform. For the Block factor, three polynomial contrast terms were used. The linear term measures constant improvement in performance over time, the quadratic term measures constant changes in the rate of improvement, and the cubic term in this case measures differences in improvements after odd vs. even blocks.

Same-different performance was a significant positive predictor of grammar-learning performance ($F(1, 68) = 4.06, p < 0.05$). The Block × Context interaction was not significant ($F(6, 207) = 0.89, p = 0.50$). Only the linear Block term was significant ($F(1, 207) = 5.20, p < 0.03$); neither the quadratic nor cubic main effects approached significance (all $p$s > 0.1). The Smoothness contrast was not significant ($F(1, 68) = 0.02, p = 0.89$), nor was the Repetition contrast ($F(1, 68) = 1.25, p = 0.27$).

6.3.1.2 Trimmed Sample

A different pattern of results emerges when the highest- and lowest-performing groups of participants are excluded. As in Experiment 1, when the analysis is restricted to intermediate performers, same-different performance is no longer a significant predictor of grammar-learning performance ($F(1, 32) = 0.60, p = 0.45$). The Block × Context interaction was not significant ($F(6, 99) = 1.79, p = 0.11$). The overall effect of Block was nonsignificant as well ($F(1, 99) = 1.47, p = 0.23$).
Figure 6.3: Mean Discrimination Score by Context and Block. Error bars denote ±1 SE.

Of key interest, the main effect contrast between the Smooth group and the other groups was significant ($F(1, 32) = 4.97, p < 0.04$), with superior performance in the Smooth condition. The contrast between the Repetition and Uniform groups was not significant ($F(1, 32) = 2.53, p = 0.12$). Means and standard errors are displayed in Figure 6.2b.

6.3.2 Discrimination Scores

6.3.2.1 Full Sample

The same analyses were repeated for the confidence-weighted scores. Same-different performance was a marginally significant positive predictor of grammar-learning performance ($F(1, 68) = 3.75, p < 0.06$). The Block × Context interaction was not significant ($F(6, 207) = 1.42, p = 0.21$). The linear Block term was significant ($F(1, 207) = 9.20, p < 0.005$), but both the quadratic and cubic terms were nonsignificant ($ps > 0.1$). The Smoothness contrast was not significant ($F(1, 68) = 0.09$,
\( p = 0.76 \), nor was the Repetition contrast \((F(1, 68) = 0.78, p = 0.38)\). Means and standard errors are displayed in Figure 6.3a.

### 6.3.2.2 Trimmed Sample

When the analysis was restricted to participants between the 2\textsuperscript{nd} and 3\textsuperscript{rd} quartiles of their respective context condition, same-different scores were no longer a significant predictor of grammar-learning \((F(1, 32) = 0.05, p = 0.82)\). The Block × Context interaction was not significant \((F(6, 99) = 1.56, p = 0.17)\). The linear Block term was nonsignificant \((F(1, 99) = 1.64, p = 0.20)\), as was the quadratic term \((F(1, 99) = 0.26)\). The cubic Block main effect was significant, however \((F(1, 99) = 5.74, p < 0.02)\), with better performance in odd-numbered blocks.

Of greatest interest, the Smooth contrast is marginally significant \((F(1, 32) = 3.65, p < 0.07)\), with better performance on average in the Smooth condition. The contrast between the Repetition and Uniform groups is not significant \((F(1, 32) = 1.16, p = 0.29)\). Means and standard errors are displayed in Figure 6.3b.

### 6.3.3 Comparison to Experiment 1

An overarching hypothesis is that alternating between a discrete representation of tones and a continuum representation of pitch is adaptive for learning different classes of generalization. In line with this idea, we predicted a reversal in the pattern of performance across context conditions between Experiment 1 and Experiment 2. While the evidence for interference by the Smooth context in Experiment 1 was strong, evidence for the opposite in the present experiment is somewhat weaker, with the predicted differences appearing only in some analyses. As such, a final set of analyses combined the data from Experiment 2 with a subset of the data from Experiment 1 in order to directly test for an interaction between Context and Grammatical Structure. Since only the High Variance variations of the Repetition and Smooth distributions were used in Experiment 2, the two Low Variance groups in Experiment 1 were not included in the combined analysis.
As with the individual analyses, both full sample and trimmed sample analyses were conducted.

### 6.3.3.1 Number Correct

A Test Grammar (Repetition vs. Contour) × Context (Smooth, Repetition and Uniform) × Block (1-4) + Same Different Score ANCOVA was performed on the numbers of correct responses in each block for the full sample. Same-different scores were a significant positive predictor of grammar-learning ($F(1, 137) = 13.42, p < 0.0005$). The three-way interaction was not significant ($F(6, 414) = 0.54, p = 0.77$).

Of greatest interest, the interaction of Grammar with the Smooth contrast was significant ($F(1, 137) = 4.91, p < 0.03$). The Grammar × Repetition contrast was not significant ($F(1, 137) = 0.46, p = 0.50$).

The interaction between Grammar and the linear Block term was significant ($F(1, 414) = 18.61, p < 10^{-4}$), with a greater slope over blocks for the grammar in Experiment 1. Neither the quadratic nor cubic components differed significantly between grammars ($ps > 0.2$).

The main effect of Grammar was significant ($F(1, 137) = 18.89, p < 0.0001$), with superior performance in Experiment 1. The mean total numbers of correct responses, summed over blocks, are displayed in Figure 6.4a, broken down by Experiment and Context condition.

When the analysis was restricted to the trimmed sample, same-different scores were no longer a significant predictor ($F(1, 65) = 2.44, p = 0.12$). The three-way interaction between Grammar, the Smooth contrast, and the quadratic Block term was significant ($F(6, 198) = 2.30, p < 0.04$). This interaction is driven by the Smooth contrast and the quadratic Block term ($F(1, 198) = 9.00, p < 0.005$). In Experiment 1, it was the Repetition and Uniform conditions which exhibited greater concavity, whereas in Experiment 2 the Smooth condition’s mean learning curve had the greater downward bend. None of the other three-way contrasts were significant ($ps > 0.2$).

Most importantly, the two-way interaction between Grammar and the Smooth contrast was significant ($F(1, 65) = 18.40, p < 0.0001$), reflecting a meaningful reversal in the effect of Smoothness
The interaction between Grammar and the linear Block term was significant ($F(1, 198) = 29.40, p < 10^{-6}$), with a greater slope over blocks for the grammar in Experiment 1. Neither the quadratic nor cubic components differed significantly between grammars ($p_s > 0.15$).

The main effect of Grammar was significant ($F(1, 65) = 39.54, p < 0.0001$), with superior performance overall in Experiment 1. Means and standard errors, summing over blocks, are displayed in Figure 6.4b.

### 6.3.3.2 Discrimination Scores

When confidence-weighted scores were analyzed for the full sample, same-different scores again appeared as a positive predictor of performance ($F(1, 137) = 14.25, p < 0.0005$). The three-way interaction was non-significant ($F(6, 414) = 0.88, p = 0.51$).
Figure 6.5: Mean Overall Discrimination Score by Experiment and Context. Error bars denote ±1 SE.

Most importantly, the interaction between Grammar and the Smooth contrast was significant \((F(1, 137) = 4.02, p < 0.05)\), but the interaction between Grammar and the Repetition contrast was not \((F(1, 137) = 0.08, p = 0.78)\).

The interaction between Grammar and the linear Block term was significant \((F(1, 414) = 18.45, p < 0.0001)\), with greater overall learning slopes for the Repetition grammar, but the interaction of Grammar with the quadratic Block term was nonsignificant \((F(1, 414) = 0.02, p = 0.88)\), as was the interaction between Grammar and the cubic Block term \((F(1, 414) = 0.42, p = 0.52)\).

The main effect of Grammar was significant \((F(1, 137) = 18.10, p < 0.00001)\), with superior overall performance in Experiment 1. Means and standard errors, averaging over blocks, are displayed in Figure 6.5a.

When the analysis was restricted to the trimmed sample, same-different scores were a marginally significant positive predictor of grammar-learning \((F(1, 65) = 3.01, p < 0.09)\). The three-way
interaction was significant overall ($F(6, 198) = 2.31, p < 0.05$), with this effect being driven by the interaction between Grammar, the Smooth Context contrast, and the quadratic Block term ($F(1, 198) = 11.05, p < 0.005$). This reflects the fact that the Repetition and Uniform context condition exhibited greater concavity in their mean learning curves in Experiment 1, whereas the Smooth context condition exhibited the greater bend in Experiment 2. None of the other three-way contrasts were significant ($p$s $> 0.2$).

Of greatest interest, the interaction between Experiment and the Smooth Context contrast was significant ($F(1, 65) = 15.08, p < 0.0005$),反映 the reversal in the effect of Smoothness between experiments. The Grammar $\times$ Repetition contrast was not significant ($F(1, 65) = 0.13, p = 0.72$).

The Grammar $\times$ Block interaction was significant overall ($F(3, 198) = 10.87, p < 10^{-5}$), with greater positive slope in Experiment 1 ($F(1, 198) = 30.20, p < 10^{-6}$). The other components of Block did not interact significantly with Experiment.

The main effect of Grammar was significant ($F(1, 65) = 31.14, p < 0.0001$), with superior performance in Experiment 1. Means and standard errors, averaging over blocks, are displayed in Figure 6.5b.

### 6.4 Discussion

The present experiment was closely parallel to Experiment 1, with one key change: instead of a rule in which discrete repetition distinguished grammatical from ungrammatical sentences, here the rule relied on pitch contour, a property which draws on the continuous dimension of frequency that underlies the tone vocabulary. As before, the key manipulation within the experiment was the presence or absence of a “smoothness constraint”, which imposes a gradient pressure for intervals between successive tones to be small in magnitude. In Experiment 1 the presence of such a constraint was hypothesized to explain away frequent repetitions, reducing their salience and leading to greater difficulty in learning a rule that depended on their presence. The motivation for Experiment 2 was the hypothesis that, in addition to explaining away repetitions, the smoothness constraint would
highlight the underlying pitch continuum, thereby increasing the prior plausibility of the contour rule.

This prediction was generally borne out, particularly in the analyses that focused on intermediate performers: In a stark reversal from Experiment 1, learners in the Smooth context condition were better able to learn the contour rule than were learners in the Repetition and Uniform conditions. As before, these two conditions were not significantly different from each other, which here is not surprising, since neither obviously offers better support for the relevance of contour.

Although the planned analysis for the trimmed sample revealed the expected advantage for the Smooth group compared to the other two groups, and although the Repetition and Uniform groups did not significantly differ from each other, if one did not have those contrasts in mind in advance, merely inspecting the means across the three groups would not produce the impression that the Smooth group stood out from the other two. This is mainly due to the odd pattern of performance in the Repetition condition, which exhibited sharp improvements and declines in performance from block to block, doing nearly as well as the Smooth group in Block 3. This volatility is difficult to explain, although it resembles somewhat the block effects seen in the 4-month-olds who had been familiarized with the Repetition context in Experiment 2. Conceivably, a similar explanation could apply here: Learners in the Repetition condition initially entertain multiple hypotheses about the nature of the interval distribution, but come to settle on the one that is most consistent with the distinction being made in the test items. Although it is not clear why, if this is the case, a similar discontinuity was not seen in Experiment 1, which shares its population with this experiment and its grammar with Experiment 2, one might speculate that, for the infants in Experiment 2, the hypothesis that is preferred initially is the Smooth hypothesis, whereas for adults in Experiments 1 and 3 it is the Repetition hypothesis; thus in Experiment 1, learners' initial guess is correct. This difference could be due either to \textit{a priori} differences in the two populations, or to differences in the amount of exposure to the Repetition distribution. After all, the adults in Experiments 1 and 3 heard twice as many context melodies as did the infants in Experiment 2, and hence they could
be expected to more strongly suppress a Smoothness hypothesis after exposure to the Repetition distribution.

It is worth commenting also on the oddly shaped learning curves that appear in this experiment. Unlike in Experiment 1, where, on average, steady improvement over blocks was exhibited in all conditions, here, mean performance declined at one point or another in each context condition. In some analyses, the cubic component of the Block effect was significant, suggesting that learners had an easier time distinguishing grammatical from ungrammatical sequences in test blocks where the melodies were instantiated in the same tone vocabulary as in the training blocks. Together, these learning curve characteristics are suggestive of a less explicit learning style than that in Experiment 1, where participants who performed above chance were usually able to articulate the grammar rule. In contrast, few participants in the present experiment were able to describe the contour rule explicitly, even while many had correct answer rates statistically above chance.

There is at least one alternative hypothesis that is consistent with both a Smoothness advantage in this experiment and a Smoothness disadvantage in Experiment 1. It is possible, perhaps even likely, that the Smooth context melodies sounded more like music than those in the other two conditions. This could bias participants to look for structure that is consistent with the music they are familiar with, which may turn them away from rules of the sort used in Experiment 1, and toward the sort of contour rule used here. While this account would still mean that learners are encoding the Smoothness property, the connection between this property and the type of structure that is easy to learn may be less direct than I have proposed.

One piece of evidence that may count against this alternative is that the contour rules appeared to be more difficult to learn overall than the repetition rules. If the mere fact of hearing music biases learners against discrete repetition rules, one might expect the task in Experiment 1 to be more difficult overall. However, the difficulty in the present experiment could simply reflect greater complexity: in order to distinguish grammatical from ungrammatical melodies in the repetition case, learners could simply focus on the first two or last two notes, whereas to distinguish “arc”
from “zig-zag” melodies, at least three pitches had to be taken into account. Moreover, to learn the contour grammars used in the present experiment, learners were required to abstract across inverted contours, lumping together upward and downward arcs, and similarly combining rise-initial and fall-initial zig-zags. Due to these inherent differences in the level of complexity, no conclusions can be made from existing data about learners’ prior weights over the kinds of structure likely to be present in music. However, we have also seen in Chapter 5 that 4-month-olds, who appear unlikely to encode natural music as Smooth, nonetheless exhibit the effect expected by an explaining away hypothesis. For at least this group, it would appear unlikely that prior biases drive the observed effect. It would, nonetheless, be interesting to direct future research toward this issue of disentangling prior biases from experiment-internal learning.

A natural follow-up to Experiment 3 would be to simplify the contour grammar by asking participants to learn only one contour, rather than collapsing across two. In addition to better lending itself to inferences about prior biases, an experiment which is matched in overall difficulty to Experiment 1 would simplify the interpretation of the Grammar × Context interaction, and would rule out the possibility that the failure to see differences between conditions is simply due to a floor effect.

Another manipulation that would get at the biases that learners bring in to the experiment concerns the nature of the instructions. In both of the experiments described here, the tone sequences being heard were described as “sentences” in an alien “language”. It is possible that hearing the input described in this way causes learners to engage their prior beliefs about the kinds of structure found in language, despite the fact that, on the surface, the sequences are more similar to musical melodies (albeit with a strange tonality) than they are to linguistic utterances. If, instead, the sequences were described up front as melodies, it is possible that the pitch continuum would be more salient a priori, thereby improving performance in the contour task relative to the repetition task. It would also be informative to examine the explaining away phenomenon in a novel domain, so as to tease apart prior biases from learning that takes place in the laboratory.
CHAPTER 7

GENERAL DISCUSSION AND FUTURE DIRECTIONS

7.1 Summary and Conclusions

In three experiments, it was shown that the statistical information present in an artificial musical context has an influence on how difficult it is for both adults and infants to learn an abstract grammatical rule. In Experiments 1 and 2, a general tendency for the intervals between successive pitches to be small in magnitude — what I have termed a "Smoothness constraint" — interferes with the ability of both adult and infant learners to extract a positional repetition rule from a set of examples. Experiment 3 showed that the Smoothness constraint does not interfere with rule-learning generally, and indeed it can facilitate learning, provided the rule being learned is consistent with the higher-level mode of organization reflected by the constraint. The model described in Chapter 4 shows that the pattern of results observed, in at least Experiment 1, is predicted by an “ideal observer” which entertains two qualitatively distinct processes for generating sequences of tones, and infers their relative prominence, along with their parametric settings, from data in accordance with Bayes’ Rule.

A specific implication of the present findings concerns the question of domain-specificity in rule-learning. The three experiments reported here show that rule-learning can become more or less difficult following even relatively brief experiences in the laboratory. It stands to reason, then, that much greater quantities of experience with the natural environment will affect rule-learning abilities as well, potentially creating discrepancies among domains (such as language and music). These discrepancies do indeed reflect a form of domain-specific constraint on learning; however, as the present studies have shown, such constraints can emerge on a developmental, rather than
evolutionary, time scale.

More generally, I have suggested that the overall pattern of results across all three experiments constitutes evidence that, at least in open-ended circumstances with few specific task demands, humans are “abductive”, explanatory learners. That is, they make inferences about aspects of the environment that can not be observed directly by constructing and testing hypotheses, which allow them to simulate a range of possible data and assess how well their observations fit with each of multiple possible explanations. This picture of learning fits with the “Theory Theory” in developmental cognitive psychology (Gopnik, 2000), which contends that young learners in particular reason (explicitly or implicitly) about the hidden mechanisms underlying their world.

7.2 Questions, Limitations and Ambiguities

While the present results are promising for a theory of adaptive and abductive learning, they leave a number of interesting questions unanswered. I will first discuss several unanticipated patterns in the data that are suggestive of additional complexity in the specific forms of learning occurring during the experiments. I will then move on to broader issues.

7.2.1 The Repetition Distribution

One curious aspect of all three experiments concerns the contribution of repetition rate toward grammar-learning performance. In Experiment 1, the two Repetition groups clustered much more closely with the Uniform group than with the Smooth groups, despite having repetition rates that were matched to a same-variance Smooth counterpart. This shows that, at a minimum, more of the interval distribution than just the repetition rate is taken into account by learners faced with learning a repetition rule. However, although repetition rate did not significantly impact performance, the numerical rank in performance across conditions suggests that it may be playing a small role on top of the larger one played by the Smoothness constraint. Interestingly, at least for the adult learners, the effect of added repetitions was facilitative, if it existed at all: In every
pair of conditions that did not differ in Smoothness (i.e., both pairs of Low and High Variance groups, as well as the Repetition and Uniform groups), the group exposed to more repetitions performed better. Rather than desensitizing learners, the extra repetition may have highlighted the fact that it was an important feature of the environment. This is consistent with the notion that part of learning is settling on a useful representation for one’s input. In addition to determining useful correlations and predictive relationships among features of the environment, a learner must determine what sets of features to learn about in the first place.

The performance of the Repetition groups in Experiments 2 and 3 is more difficult to explain. The 4-month-olds in the Repetition context in Experiment 2 appeared more similar to those in the Smooth condition than to those in the Uniform condition when looking times were collapsed across Blocks, contrary to prediction. When the last block of test trials is isolated, however, the pattern of looking times is as predicted. I have speculated that the intermediate and discontinuous pattern of preferences in the Repetition condition may reflect a sort of “ambidistributionality” of the interval pattern in that context (to coin a somewhat ridiculous term). That is, while the intended character of the Repetition distribution was as a uniform distribution with an added specific feature (namely lots of repeated tones), if one has a strong enough prior expectation that melodies are Smooth, one can force the Repetition distribution into the Smooth model with greater success than one would have with the Uniform model (provided a distribution with an extremely large variance in intervals is not considered a good exemplar of “Smooth”). As such, it is possible that the 4-month-olds simultaneously maintained two hypotheses pertaining to the underlying musical properties that gave rise to the Repetition distribution, but settled on the “uniform plus repetitions” interpretation after hearing the contrast between AAB and ABB at test.

A similar explanation might apply to the strange pattern of results in the Repetition condition in Experiment 3. There, as in the preceding experiments, the original prediction was that the Repetition group would cluster with the Uniform group, though here the direction of performance was expected to be reversed due to the use of a “continuum-based” grammar. However, as with
the 4-month-olds in Experiment 2, participants in the Repetition condition were solidly between the other two groups overall. Also similar to the 4-month-olds, this group of adult learners showed a sharp improvement in performance part way through the testing phase. This pattern of results is again consistent with learners maintaining multiple hypotheses about the nature of the mechanism generating the interval distribution, and making use of the test contrast to resolve that ambiguity. Unlike the 4-month-olds, however, the adults in Experiments 1 and 3 appear to initially classify the Repetition context as reflecting a non-Smooth environment, with learners in Experiment 3 revisiting that assumption upon encountering new information consistent with the notion that the underlying pitch continuum really is important in this environment. It is unclear why 4-month-olds should prefer the continuous interpretation initially, while the adults prefer the discrete one, though as discussed in the conclusion to Chapter 6, this may be a function of the different quantities of exposure involved: After hearing 200 melodies (rather than the 100 heard by the infants), adult learners might be expected to be more likely to arrive at the correct (non-Smooth) interpretation. The instructions given to the adults may also play a role here: The adults were told that they were learning a language, which may well have biased them toward a discrete interpretation.

One more reason why the Repetition condition might alternately facilitate and interfere with learning may be more naturally integrated with the abductive machinery that drives explaining away. It is instructive to consider the dual role of the information presented in the Context phase. On the one hand, these sequences provide information about the abstract nature of the environment: To what extent is the underlying pitch continuum relevant in how melodies are organized? However, the context sequences can also be supposed to contain some grammatical melodies, even though it is unclear which these are.

In the case of the Smooth condition, these two kinds of information work in the same direction, namely against a repetition rule and in favor of a contour rule. First, the prevalence of small intervals suggests that the pitch continuum is important, which it turns out not to be for the rule being learned. Moreover, the high rate of repetitions in all positions in the context melodies may
make it harder to notice the positional regularity present in the grammatical melodies. Nothing about the specific melodies heard should be expected to inform any particular contour rule, since the contours are completely random, though a case could be made that a more repetitive environment would contribute less noise to learning a contour rule – after all, there are fewer instances of intervals moving in the wrong direction.

In the Repetition condition, it is less clear how the context informs the abstract nature of the environment. On the one hand, if the strongest \textit{a priori} alternatives to a continuum organization feature same/different relationships (as is the case in the Bayesian model discussed in Chapter 4), then the Repetition condition should provide even more evidence for this organization than the Uniform condition. On the other hand, if, rather than choosing between two alternatives, one is deciding in favor of or against the continuum hypothesis, the Repetition distribution arguably provides less evidence against it than the Uniform distribution, as discussed in the introduction to this chapter. In either case, however, the extra repetitions across context melodies may obscure the positional regularity heard later. Whereas the model in Chapter 4 treats positions as independent, humans may expect it to be unlikely that a repetition rule depends on position if their prior experience indicates that repetitions are evenly distributed across positions. In contrast, as discussed above, repetitions could reduce the noise contributing to learning a contour rule.

In any case, there are clearly more potential moving parts than there is data at present, and so there is plenty of room for further research, both behavioral and computational, into each of the possibilities discussed here.

\textbf{7.2.2 By What Path Does Smoothness Influence Rule-Learning?}

As pointed out earlier, a possible confound complicating the interpretation of the Smoothness effect is the possibility that Smooth melodies sound more like music, and hence may engage prior biases about structure found in melodies that have nothing to do with explaining away repetitions. While I have argued here that Smoothness provides (abductive) evidence for organization of tones along
an underlying pitch continuum, thereby decreasing the evidentiary value of repetitions for a rule and increasing the plausibility of a contour rule, it is possible that the path is less direct. To wit, learners could come to the same conclusions if Smoothness is (inductively) associated with music, and music is associated with contour (and not repetition) rules. While this latter process could become internalized as a result of the former, the source of prior biases in music is precisely what is at issue, and hence it would be circular to suggest that evidence for prior biases constituted evidence for the preferred explanation.

To the extent that 4-month-olds’ performance in Dawson and Gerken (2009) indicates that they do not encode music as Smooth, it would seem difficult to attribute the context effects seen at that age in Experiment 2 to prior biases. For the other populations, however, none of the existing data can convincingly rule out the possibility of this less direct link between Smoothness and the difficulty of learning various rules in tone sequences. What makes this confound especially difficult to resolve is that, since natural melodies are smooth, the explaining away account predicts that people should develop biases of this kind, which could quite reasonably be expected to influence behavior in an experimental setting. Hence, an attempt to disambiguate these possibilities by ruling out the existence of such biases is unlikely to succeed. Similarly, arguing that Smooth melodies do not sound any more like music than their less smooth counterparts would be difficult.

This is a similar problem faced when studying language learning: Using linguistically natural inputs, while conferring a certain ecological validity on an experiment, brings with it the same confound between prior experience and in-laboratory experience. Indeed, we cited precisely this issue in Dawson and Gerken (2009) in cautioning against hasty inferences about the source of apparently domain-specific rule-learning biases. In studying language learning, many researchers employ artificial grammars in an effort to separate prior experience from in-laboratory learning. Perhaps extending this approach to higher-order forms of learning will prove fruitful. Although it is not possible to take prior experience out of the equation entirely, if learning can be shown to exhibit similar abductive dynamics in analogous situations across multiple domains, spanning the
continuum between ‘natural’ and ‘artificial’, we will have converging evidence that a major role of learning is causal explanation.

7.2.3 Conditional vs. Marginal Interval Distributions

In all three experiments, the Context melodies were constructed according using one of three distributions. In the Uniform condition, following any given tone, each of the remaining tones was equally likely. In the Smooth conditions, the probability of a tone fell off gradually as the number of steps increased. Finally, in the Repetition conditions, the same tone was likely to be repeated, but all other tones were equally likely. Although the interval distributions following any particular tone could be simply described in this way, the marginal interval distributions were more complex due to the restricted range: the largest intervals were only possible following a tone at one of the extremes of the range (see Fig. 3.2). As a result of this restricted range, the overall interval distribution in every condition contained more small intervals than large ones. It is possible that if participants were not sensitive to the bounded pitch range, they may have encoded some degree of Smoothness even in those conditions where it was intended to be absent. This could have obscured interesting differences across conditions, and perhaps contributed to the somewhat messy pattern of results in Experiments 2 and 3.

Another avenue for exploration might be to manipulate the Context melodies so as to achieve flat marginal interval distributions in the non-Smooth conditions. Although this would necessitate conditional distributions with an apparent “anti-Smoothness” constraint, it is certainly plausible that listeners do not keep separate statistics for each preceding tone.

7.2.4 Blocking and Generalization Over Vocabularies

In the two experiments with adult participants (Exps. 1 and 3), the Grammar-Learning phase was divided into four blocks, each with a training component and a test component. In odd-numbered test blocks, the probe items were instantiated in the same tone vocabulary as the training
items. In even-numbered blocks, the test melodies were instantiated in a novel vocabulary. The motivation to include both types of test items was the added ability to explore differences in the degree of abstraction of the learned generalization. It was supposed that differences in performance across context groups could be more pronounced when an abstract (i.e., vocabulary-independent) generalization was required in order to successfully discriminate grammatical from ungrammatical test items, as more abstract forms of learning might be expected to be more model-dependent.

In the two adult experiments discussed here, no significant \( \text{Context} \times \text{Block} \) interaction was found in any analyses, and hence, although the change in vocabulary may have affected learning overall (and in different ways depending on the grammar being learned), there is no evidence of a differential effect of context (and, by hypothesis, a domain model) on more or less abstract generalizations. As such it may prove simpler in future experiments to use only one vocabulary throughout the grammar-learning phase, or to avoid alternation between training and test. Although the latter would come at the expense of the ability to detect improvement in performance over blocks, it would reduce the “backward” influence of exposure to test items on learning during training.

7.3 Broader Open Questions

While it is undoubtedly fascinating to think about each detail of the preceding experiments in excruciating detail, the data they yield is only useful if it contributes to understanding broader principles at play in learning. While the interpretation of the three experiments presented here is far from unambiguous, I believe it does at least hint at a greater role for explanatory, abductive learning than previously supposed.

7.3.1 The Role of Surprise

Recall C.S. Peirce’s (1935) description of abduction:

The form of the inference \ldots is this:
1. The surprising fact, C, is observed.

2. But if A were true, C would be a matter of course.

3. Hence, there is reason to suspect that A is true.

One of the primary virtues of an explanation-based framework of learning is that the learner can monitor discrepancies between her input and her expectations; that is, surprise. The task of probabilistic abduction can be described, in a nutshell, as surprise-reduction: shift “belief” to hypotheses to the extent that they make the data less surprising. In the experiments discussed in this dissertation, a learner with a Smooth model of melodies should find repetitions less surprising than a learner without that model, and hence the adoption of a repetition hypothesis does less work.

The ideal probabilistic abductive learner can monitor not only the relative degree of surprise resulting from accepting one hypothesis compared to another, but can also track the absolute degree of surprise. The latter quantity constitutes a potential signal to trigger the formation of new hypotheses: if one’s input is highly surprising, aggregating over all currently available hypotheses, this is a quantifiable motivation to generate new hypotheses. Although the way in which these new hypotheses might be generated, as well as their scope, remains a major puzzle for future research, surprise (particularly as quantified by aggregate likelihood under a set of hypotheses) provides a principled criterion for determining when the current hypothesis space must be expanded and/or differentiated.

1Ultimately, any satisfying theory of hypothesis-generation will likely need to be worked out across levels of abstraction. Although a formal mathematical level of description has no problem positing a space of infinite potential hypotheses, cognition takes place in finite brains, and so actual hypotheses must emerge from the Platonic world of potential infinity to be instantiated in a neural representation. As such, an important avenue of research will be to formalize the connections between Bayesian probabilistic models, at a computational level of abstraction, and the algorithmic models of connectionist networks; regarding which, see the discussion just below.
The Connectionism Connection

At the beginning of this dissertation, I contrasted inductive forms of learning, in which the outcome of learning is a set of predictive relationships, with abductive learning, in which explanatory models are employed to make inferences about the underlying causes of one’s observations. It was suggested that feed-forward connectionist networks can be described as devices for (probabilistic) induction, whereas graphical model formalisms from the field of machine learning are idealized implementations for theory-based learning. While these two forms of inference have been presented as alternatives to each other, the implementations discussed represent descriptions at rather different levels of abstraction. In the terms of Marr (1982), the multilayer perceptrons of Rumelhart and McClelland (1986) work out the induction problem at an algorithmic level. Moreover, to the extent that these networks are informed by principles of neural computation, they represent an attempt to approximate aspects of the implementational level of biological cognition as well. In contrast, the graphical model framework represents a computational level of description, specifying only what quantities are to be computed, and not how to compute them.

If the Theory Theory is to become a fully developed account of learning and inference, it will be necessary to work out whether appropriate brain mechanisms exist to carry out the computations required to manipulate high-dimensional probability distributions in real time. Fortunately, some of these details are beginning to be worked out (see Doya, Ishii, Pouget, & Rao, 2007, for a review). R. P. N. Rao (2004) demonstrated that a connectionist architecture featuring recurrent connections — that is, where activity in neurons closer to the input signal can be influenced by activity “later” in processing, as well as the other way around — can implement near-optimal probabilistic inference over a hidden Markov model. Litvak and Ullman (2009) developed a simulated network of neurons which rapidly estimates the parameters of arbitrary probabilistic graphical models using algorithms related to the Belief Propagation algorithm developed in machine learning (e.g., Pearl, 1988). Here, too, the presence of recurrent connections is a key component of achieving optimal inference.
By allowing information to flow in both directions, these networks naturally strike a balance between bottom-up (data-driven) and top-down (expectation-driven) information. In Bayesian terms, these two sources of information are roughly analogous to the likelihood and the prior terms that inform the posterior distribution over the unobserved quantity or category which is relevant to behavior. Explaining-away behavior could potentially emerge as well, provided feedback connections are allowed to be inhibitory, effectively suppressing predictable information. Exactly this manner of suppression has been proposed in ‘efficient-coding’ and ‘predictive-coding’ theories of neural computation (Mumford, 1992; R. Rao & Ballard, 1997), and is supported, at least in visual cortex, by functional imaging during shape perception (e.g., Murray, Kersten, Olshausen, Schrater, & Woods, 2002). If the neural substrates underlying more abstract inferences — about language, music, high-level causal hypotheses, etc. — can be shown to exhibit similar dynamics, we will be on our way to a domain-general theory of cognition.
REFERENCES


Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function,


Marcus, G. F. (1999a). Do infants learn grammar with algebra or statistics? response to seidenberg...


In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (p. 663-668).


rule learning is not specific to language. *Cognition, 105*, 669-680.


Watt, H. J. (1924). Functions of the size of interval in the songs of Schubert and of the Chippewa


