Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Mathematical Psychology 50 (2006) 339–368
Journal of Mathematical Psychology
www.elsevier.com/locate/jmp

An algebra of human concept learning

Jacob Feldman

Department of Psychology, Center for Cognitive Science, Rutgers University, New Brunswick, 152 Frelinghuysen Road, Piscataway, NJ 08854, USA

Received 8 July 2005; received in revised form 3 January 2006
Available online 2 May 2006

Abstract

An important element of learning from examples is the extraction of patterns and regularities from data. This paper investigates the structure of patterns in data defined over discrete features, i.e. features with two or more qualitatively distinct values. Any such pattern can be algebraically decomposed into a spectrum of component patterns, each of which is a simpler or more atomic "regularity." Each component regularity involves a certain number of features, referred to as its *degree*. Regularities of lower degree represent simpler or more coarse patterns in the original pattern, while regularities of higher degree represent finer or more idiosyncratic patterns. The full spectral breakdown of a pattern into component regularities of minimal degree, referred to as its *power series*, expresses the original pattern in terms of the regular rules or patterns it obeys, amounting to a kind of "theory" of the pattern. The number of regularities at various degrees necessary to represent the pattern is tabulated in its *power spectrum*, which expresses how much of a pattern's structure can be explained by regularities of various levels of complexity. A weighted mean of the pattern's spectral power gives a useful numeric summary of its overall complexity, called its *algebraic complexity*. The basic theory of algebraic decomposition is extended in several ways, including algebraic accounts of the typicality of individual objects within concepts, and estimation of the power series from noisy data. Finally some relations between these algebraic quantities and empirical data are discussed.

Keywords: Concepts; Induction; Complexity; Learning

1. Introduction

One of the most basic functions of human cognition is the discovery of patterns and regularities in data. Suspected patterns are the basis for prediction of future trends, as well as for the comprehension of past observations. Yet not all patterns seem equally plausible as the basis for prediction; and, perhaps consequently, not all patterns are equally prone to be noticed by human observers. The nature of the patterns that human observers do notice, of course, reflects our implicit assumptions about what types of regularities actually govern the data, and thus are likely to survive when new data are received.

In this spirit, much research in the 1960s was aimed at understanding what kinds of logical rules among Boolean data tend to be easily apprehended by human learners. By far the most famous finding in this connection was that conjunctive rules are more easily learned than disjunctive ones (Bruner, Goodnow, & Austin, 1960; Hovland, 1952), with a small number of other bivariate rule types following what eventually became a well-established difficulty ordering (Bourne, 1970; Haygood & Bourne, 1965). Even as early as 1961, though, a well-known result of Shepard, Hovland, and Jenkins (1961) involving trivariate concepts (discussed below) demonstrated influences of logical form that could not be understood simply in terms of a preference for conjunctions, nor indeed any other basic principles known at the time. Neisser and Weene (1962) (see also Haygood, 1963) had proposed a principle of *simplicity* or parsimony, suggesting that difficulty ordering of the concepts known at the time (not including Shepard et al.'s set of six) could be explained by the number of operations required to express them—i.e., by their complexity—but this idea was not pursued at the time. Much later, Rosch (1973) proposed "cognitive economy" as a central principle of categorization, although without any associated scheme for quantifying it. Similarly a human bias towards simple inductions was noted by Medin,

E-mail address: jacob@ruccs.rutgers.edu.

0022-2496/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2006.03.002

Wattenmaker, and Michalski (1987). More recently, studying a much wider range of Boolean concepts (all logically possible concepts involving two, three, or four examples defined over three or four features; see below for extensive discussion), Feldman (2000) corroborated the complexity effect, showing that learnability of Boolean concepts is well-predicted by their *Boolean complexity*, defined as the length (in literals, i.e. positive or negative mentions of a variable) of the shortest propositional formula equivalent to each concept. This finding suggests an intriguing connection between the subjective learnability of a concept and its degree of regularity (Feldman, 2003b).

However, traditional propositional calculus—whether using the conventional three-connective basis $(\wedge, \vee, \neg)$ or an alternative, e.g. the single-connective basis used by Haygood (1963)—makes a notoriously infelicitous choice as a language for psychological concepts. First, propositional calculus is intrinsically limited to binary (Boolean) features, while human learners are quite obviously capable of incorporating larger numbers of values per attribute (as was even recognized in the 1960s) as well as continuous-valued features, upon which much modern research has focused (e.g. see Ashby & Gott, 1988).

Perhaps more fundamentally, the expression of concepts as Boolean formulae began to appear strikingly inapt as more "graded" notions of category membership were introduced by Posner and Keele (1968) and Rosch (1973). The binary distinction between membership, defined by satisfaction of a defining logical formula, and non-membership, defined by falsification of the formula, is now universally rejected as a model of human concepts (though see Fodor, 1994)—even for those rare concepts that might properly obey such definitions (Armstrong, Gleitman, & Gleitman, 1983). Psychologically judged membership in categories is not dichotomous, but rather exhibits variation in judged typicality, ranging from highly characteristic or prototypical objects to highly uncharacteristic or peripheral ones. Propositional calculus does not lend itself to making this distinction, in part because the propositional representation is not in any obvious sense divided up into "more essential" and "less essential" elements, but rather expresses each concept in terms of a single unitary rule.

1.1. Modern accounts

Modern theories of category representation, oriented primarily towards explaining this typicality gradation, are primarily based on the notion of a similarity space (Attneave, 1950; Shepard, 1957), in which subjective dissimilarity is captured by some formal metric. Given this metric, category membership is usually characterized in terms of the dissimilarity between a given object and either an abstracted central tendency of the known examples, called a *prototype*, or—in what has become the more prominent class of theories—a larger set of stored objects, called *exemplars* (Hayes-Roth & Hayes-Roth, 1977; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1988). In exemplar theories, the observer explicitly stores many (or all) of the observed objects as examplars, along with their category labels. New objects are then classified by comparing them to the stored exemplars. Such theories, including the Generalized Context Model (GCM) (Nosofsky, 1988) and ALCOVE (Kruschke, 1992), have been very successful in explaining category membership judgments in a wide range of contexts. Yet as has been often noted, exemplar theories are striking in their complete lack of any *abstraction* from the examples; by design, such theories learn categories without (at least overtly) extracting any central tendencies or regularities from the observations at all. Qualitative patterns in the observed members of a category that one might think salient, such as that apples are usually round, or that water is usually wet, play no explicit role in exemplar theories; such patterns make their presence felt only indirectly via their presence in the exemplar set.

The rise of exemplar models has not been without controversy. Homa, Sterling, and Trepel (1981) have suggested that human observers only rely heavily on exemplars after brief exposure to the examples, and that with greater exposure prototypes form. Medin and Bettger (1994) found effects of the order of the presentation of examples that, they argued, would be hard to explain in an exemplar framework. More recently, Minda and Smith (Minda & Smith, 2001, 2002; Smith & Minda, 1998, 2000) have argued that the evidence in favor of exemplar models has generally been overestimated, and that, broadly speaking, prototype models outperform exemplar models when the playing field is leveled. One element of their argument is the historical choices of concept types to studied. Many studies have used the same few types, which may have unintentionally favored exemplar-based learning strategies, because they comprise a relatively small set of relatively heterogeneous objects (Smith & Minda, 1998). For example Minda and Smith (2002) count 30 separate uses of one concept type, originally introduced by Medin and Schaffer (1978) in support of their original exemplar model. Recently Blair and Homa (2003) have argued that this concept is so heterogeneous and phenomenally random that subjects learn it using a pure-memorization strategy, demonstrably not involving categorization at all. One immediate conclusion from this controversy is that, when evaluating the performance of competing theories, it is imperative to consider a more comprehensive range of concept types—including both regular and irregular types as well as everything in between—a goal that will be pursued below.

Heterogeneous or disjoint categories are *incompressible*, meaning they have high complexity in the sense of Kolmogorov (see Chater & Vitányi, 2003; Li & Vitányi, 1997). This makes them difficult to summarize and thus to learn (Feldman, 2000; see also Feldman, 2003b). More specifically they do not lend themselves to any learning strategy that depends on extracting regularities,

because—by definition—they do not *have* many regularities. Hence the controversy over exemplar models reflects disagreement in the literature about the role and importance of "structure" and regularity in the concepts to be learned. Exemplar models are by their very nature insensitive to structure per se; they store the examples as they come, without attempting to find any common trends. Hence the modern literature leaves substantially unresolved one of the main questions in the early artificial-rule-learning literature: the nature of the structures that human learners are predisposed to favor, such as the types of featural patterns that predispose them to group items together into categories.

Logical formalisms are indeed helpful in expressing structure, which is perhaps why early research relied on them so heavily. But again simple propositional representations of categories fail as psychological models in a number of respects, most notably that they fail to in any way distinguish between more and less typical category members. Even without this more modern desideratum, though, the inadequacy of simple propositional expressions for concepts was long clear. As mentioned, there was already evidence by 1961 that for trivariate concepts a simple preference for conjunctive structures was insufficient to explain the data. For higher dimensions (more features in play) the theoretical principles break down completely. Hence what is needed is a more sophisticated way to represent regular structure in discrete featural data, in order to (a) represent structural patterns in a psychologically revealing way, and (b) empirically test whether human learners are, in fact, sensitive to the degree of structure.

1.2. Scope and limitations

With all these factors in mind, the current article develops a new approach to representing the structure of featural patterns. We will focus on the qualitative structure of such patterns, i.e. to the set of feature vectors observed by the learner, without regard to the number of examples of each feature vector observed. (A later section extends the theory to handle variations in frequency among qualitative cases.) Although built from logic-like formalisms, the new approach aims to be more psychologically apt than propositional representations in two main respects. First, we limit ourselves to formalisms that can be extended to discrete features with arbitrary numbers of attributes per features (although for convenience we lay out the theory using conventional Boolean notation at first and only later present the generalization to n-valued features, which require more complicated notation). Second, we keep in mind that any representation of human concepts must support the distinction between more central and less central trends in the data, which in the end will make it possible for this very "logic-like" formalism to support, and in fact to accurately predict, human judgments of the degree of typicality of individual objects embedded in concepts.

Hence the main contribution of this paper is to propose a formalism for representing discrete-featured concepts that brings out the internal regularity of each concept, to the extent that it has any. Some deficiencies of the approach will be quite apparent. The theory is not presented as a "full-service" concept learning model to compete with the many extant theories that address admittedly important aspects of real categorization performance not addressed here—such as the distinction between expert and novice categorizers, the effect of task demands and prior knowledge, and neural correlates of learning processes. Rather the theory is presented to help elucidate a critical aspect of concepts defined over discrete features—their compositional structure—that is inadequately addressed in the modern literature.

One limitation of the theory deserves special mention. Unlike many contemporary theories, the proposed approach does not deal at all with continuous-valued features. Historically, and even recently, the bulk of concept learning research has used concepts defined over simple Boolean features, including such seminal modern papers as Medin and Schaffer (1978), Nosofsky (1991), and Nosofsky, Palmeri, and McKinley (1994). More recent research has begun to focus on continuous-valued features, perhaps reflecting the heavy emphasis in many current theories on similarities defined over a metric space. Theories that rely on a metric space cannot strictly speaking be applied to Boolean concepts, which lack such a metric, although in practice one can easily interpose a natural metric space between the terminal 0 and 1 corresponding to the two values of each Boolean feature. But such a metric cannot be so straightforwardly imposed on features with three or more unordered values (e.g. *shape* = `square`, `circle`, or `triangle`). Lee and Navarro (2002), attempting to apply the popular exemplar theory ALCOVE (Kruschke, 1992) to a concept set defined over two three-valued discrete features, found that ALCOVE failed to give good results unless a metric that explicitly respected the featural structure of the space was adopted. This result—while not in and of itself discrediting ALCOVE, which worked well once outfitted with feature-based metric—highlights the psychological reality and indeed centrality of featural representations, at least in some contexts. This is not to say that continuous features are not equally, or perhaps even more, important than discrete ones in other contexts; but rather that the representation of featural structure remains a critical problem, psychologically important but inadequately handled in modern theories.

1.3. Causal theories

One more motivating concern deserves special mention. Recently a number of researchers have emphasized the tendency of learners to attempt to understand observations in terms of *causal* influences among features. For example, Holland, Holyoak, Nisbett, and Thagard (1986) have

suggested that categories can be thought of as clusters of interrelated rules inferred from observations. Sloman, Love, and Ahn (1998) (see also Ahn, Kim, Lassaline, & Dennis, 2000) have argued that causal relations are essential in inducing conceptual coherence. Recently, Rehder and Hastie (2004) have formulated a model of category formation based on Bayesian belief networks, which summarize the network of probabilistic influences amid a complex joint probability distribution (Pearl, 1986, 2000).

The idea that categories originate in a network of interrelated rules has perhaps been brought out the most clearly in the idea of the "theory–theory" (Murphy & Medin, 1985). Murphy and Medin argued that psychological categories should not be thought of simply as arbitrary clusters of features, but rather in terms of the larger array of interrelated processes and causal laws in which they are embedded—the observer's "theory" of the domain or environment. Some sets of features, Murphy and Medin argued, seem to combine in a subjectively coherent way (*big scary monster*) while others do not (*colorless green idea*). Both of these feature combinations have equal status in conventional feature-based models of categories, in that each is simply an n-way combination of attributes; yet the former points to a coherent entity that resonates with a conventional understanding of the terms and their plausible modes of combination, while the latter does not. In Murphy and Medin's view, this failure to mesh with the observer's theory of the world renders the indicated concept not only unlikely, but effectively meaningless. They argued that virtually all then extant theories fail to account for how categories relate to the observer's understanding of the world.

In the ensuing two decades Murphy and Medin's argument has lost little of its force, and has been widely appreciated and frequently cited. But their point was basically negative: that conventional feature-based theories fail to account for how categories are embedded within theories. Yet building an *affirmative* theory-based account—articulating just what a theory *is*—is daunting; how do we formalize the notion of "theory", and how can we quantify the degree to which a given category resonates with it? The concept algebra is an attempt to answer this question in the narrow domain of discrete-featural patterns. The main idea will be to represent featural patterns in terms of the *implicational rules* they contain, and into which they can be decomposed. The choice of implicational rules, like the choice of conditional dependencies in Bayes nets, is motivated by the desire to pull out quasi-causal regularities and interactions the variables under observation. Indeed, the formal decomposition into quasi-causal inter-featural regularities in the concept algebra bears a close mathematical affinity to the theory of Bayes nets, as will made more clear below. Hence though it would be an overstatement to suggest that the algebraic approach described in this paper answers Murphy and Medin's call (as it certainly fails on some of their criteria for a fully theory-based account), it does draw from their arguments a critical motivation, namely the search for a more theory-like and causally oriented representation of, in this case, featural patterns.

2. Regularities and patterns

We begin by considering what sorts of patterns discrete-valued data may, in principle, exhibit. Any set of data exhibits many patterns, some of which will be simple or coarse, in the sense that they may be specified with reference to only a small number of features, while others may be complex and detailed, in the sense that they may only be described by reference to a larger number of features. Other patterns involve an intermediate number of features to specify, and thus are of intermediate complexity in this sense. One of the goals of the theory below is to capture how arbitrary patterns of data be described as *combinations of component patterns*, including simple components, complex components, and those in between. A key step is to make this complexity spectrum completely explicit, quantifying exactly how component patterns are constructed and combined. To this end, we begin by developing and classifying certain types of simple patterns in Boolean data (later extending the math to handle discrete features with larger numbers of possible values).

2.1. A motivating example

Consider an observer confronted with a "world" $\mathscr{W}$ consisting of the five amoeba-like objects shown in Fig. 1a. First, the observer must choose some "language" in which to express the structure of this world, which we think of as simply as a list $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_d\}$ of abstract property tags, called the *property language*. (For the moment we assume Boolean (binary) features, an assumption that will be relaxed later.) For the world given in Fig. 1, an appropriate language might be

$$\Sigma = \{\texttt{blob\_shaped}, \texttt{shaded}, \texttt{has\_nucleus}, \texttt{has\_dotted\_membrane}, \texttt{large}\}, \quad (1)$$

which is abbreviated to $\Sigma = \{a, b, c, d, e\}$ under the assignment

$$a = \texttt{blob\_shaped},\quad b = \texttt{shaded},\quad c = \texttt{has\_nucleus},\quad d = \texttt{has\_dotted\_membrane},\quad e = \texttt{large}. \quad (2)$$

The same world $\mathscr{W}$ encoded into Σ is shown in Fig. 1b. This is really just a set of strings, each a subset of Σ, which can be thought of as the world $\mathscr{W}$ "taken at" or "projected into" Σ, i.e. here

$$\mathscr{W}|_\Sigma = \{ab'c'd'e', abc'd'e', ab'cde', ab'cde, abcde\}. \quad (3)$$

Sometimes, when the choice of Σ is clear from context, $\mathscr{W}|_\Sigma$ will be denoted by $\mathbf{x} = \{x_1, x_2, \ldots\}$, referring to each

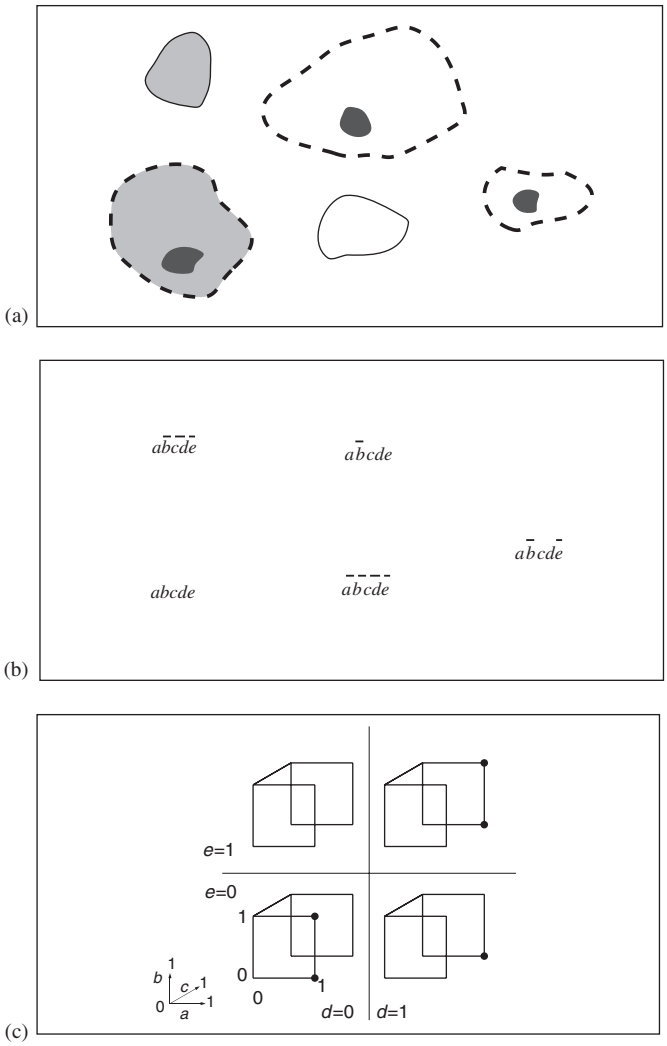


Fig. 1. (a) A "world" $\mathscr{W}$ containing amoeba-like objects. (b) The same world, encoded in the property language Σ, and (c) illustrated in abstract five-dimensional Boolean space. Heavy dots correspond to the observed examples.

individual string x_i as an *object* or *observation*. The crucial information in each object string is the set of properties that are true, and so sometimes it is convenient to think of objects as sets rather than strings, always referring to the set of "true" features.

The projection notation $\mathscr{W}|_\Sigma$ has several advantages over conventional notation in which an observation is simply a set of feature vectors, without any explicit notational acknowledgment that there is a distal world behind the features. The notation overtly recognizes the dichotomy between the abstract world $\mathscr{W}$ and the concrete set of property strings $\mathscr{W}|_\Sigma$—which only comes into existence when the $\mathscr{W}$ is "sampled" at a particular Σ.[1] There are after all in infinity of features that one might consider; this notation emphasizes that we are only considering some definite subset of them, namely those in Σ. Moreover it allows us to consider the formal question of whether the same world projected into alternative property language would or would not have analogous structure—a critical question in considering the stability of our conclusions over changes in observational conditions. Such questions could not be considered in a clean way if the *world* were identified with the *set of observations*, as in conventional feature-vector formalisms.

The set of observations $\mathscr{W}|_\Sigma$ may contain some pattern of property covariation—certain properties systematically appear or fail to appear in the company of certain others. Such patterns are of course not arbitrary, but reflect (in some possibly indirect way) the actual structure of $\mathscr{W}$—that is, the "true" pattern of causal interaction among the distal properties represented proximally by the various σ_i. The goal of the observer is to infer these distal patterns and regularities from the extremely impoverished evidence available among the observable patterns (see Barlow, 1994; Richards & Bobick, 1988; Shepard, 1989).

2.2. Logical notation

For mathematical convenience, we will adopt a logical language for expressing regularities among features. It should be understood though that this formalism does not in any way commit us to a "deductive" style of inference—indeed, the *induction* of patterns from featural data is the entire goal of the enterprise.

First, a few definitions. A *Boolean formula* is an expression built from variables (property tags) combined using the operators $\wedge, \vee$ and $\neg$ (and, or, and not) in the familiar way. Following a common convention, we will, however, adopt a more concise notation and usually write ab for the conjunction $a \wedge b$, $a + b$ for the disjunction $a \vee b$, and a' for the negation $\neg a$.

The set of objects $\mathscr{W}|_\Sigma$, regarded above as a set of strings, can also be thought of as a Boolean logical expression. For example the object set given above (Eq. (3)) is really just the disjunction

$$ab'c'd'e' + abc'd'e' + ab'cde' + ab'cde + abcde, \tag{4}$$

in which each string (object) is now understood as a conjunctive formula expressing the features of that object. This observation formula, like all sets of objects, is technically said to be in "disjunctive normal form" (DNF), meaning simply that it is a disjunction of conjunctions. The set of observations expressed as a DNF formula is a logical "theory" that is satisfied by the environment—that is, the values that the variables take on in the environment makes the expression true.

This DNF formula expressing the observations, like all Boolean propositions, is logically equivalent to an infinite number of other propositions; and similarly it *entails* an infinite number of other propositions. Such entailed propositions may be thought of as *rules* or *patterns* that

[1] What we observe is not nature itself, but nature exposed to our method of questioning. Werner Heisenberg.

hold in the world under observation. The inductive question now becomes: which entailed propositions should we believe? Which will hold in future objects generated from a similar source?

A bit of a paradox presents itself here, related to what is sometimes called the *bias-variance tradeoff* (see Hastie, Tibshirani, & Friedman, 2001; Mitchell, 1982). There is a fundamental tradeoff in induction between expressivity and predictive power. Generalization based on a very expressive representation of the data can, counterintuitively, *reduce* inferential leverage, because it entails predictions based on random characteristics of the sample set along with stable ones—often called *overfitting the data*. Propositional formulae are general enough to express *any* combination of observations in Boolean data, which means they make a poor basis for generalization. Because it can fully fit the sample set, it fits both "signal" and "noise," making it impossible to distinguish those patterns that are likely to recur (signal) from those that are not (noise). To properly express the distinction between "signal" and "noise" in featural patterns, though, we need a different pattern language.

3. An algebra of simple concepts

It is natural, then, to try to use a simpler or more expressively impoverished language in which to describe patterns. The idea is to pick out a certain atomic set of rules to use as the building blocks of our "concepts" or "patterns." These building blocks together with the rules for combining them will constitute our description language, the "concept algebra." Ideally, we seek a set of building blocks that is sufficiently rich to express some interesting concepts, but simple enough that observing them actually means something.

Consider two very simple kinds of rules that observed objects may obey.

3.1. Implication

One very obvious kind of rule—implicit in the notion of "causality"—is when one property's value determines another's. This distal causality will be reflected proximally in a pattern among observed variables, namely a logical implication:

$$\sigma_1 \rightarrow \sigma_2, \tag{5}$$

meaning that σ_1 is never true unless σ_2 is also.[2] Note that the implication here is one that holds contingently in the data, not necessarily one that holds analytically.

[2]It is useful to keep in mind that we could replace the hard-and-fast rule $\sigma_1 \rightarrow \sigma_2$ with its "soft" stochastic counterpart $p(\sigma_2|\sigma_1) > \theta$ (for some threshold θ) without affecting the algebraic structure of the theory to follow (with one technical exception noted below). In what follows, for clarity we will generally focus on the algebraic combinations of "perfect" regularities based on patterns in noiseless data. The more realistic case of estimation of imperfect regularities from noisy data is addressed in a separate section below.

Why is this particular type of rule important? Loosely, the pairwise implication is important in that it realizes the simplest possible conception of "causality" in tangible, observable form. Of course, an observer cannot infer directly from a featural pattern whether a distal relationship is, in fact, causal; such a judgment is a best provisional, and indeed exactly what is meant by "in fact causal" is controversial (see Glymour, 2002; Michotte, 1946/1963; Pearl, 2000; Salmon, 1998; Suppes, 1970 for extensive discussion). Nonetheless the observer's tacit assumption that the world is causally coherent can be connected to an expectation that observable pairwise implications tend to exist and to hold consistently throughout $\mathscr{W}$; and individual such implications serve, at least, as rough hypotheses about the identity of those relationships.

3.2. Affirmation

An even simpler type of consistent world structure is when one property σ holds consistently in all observed objects, i.e. a rule with the very simple form

$$\sigma. \tag{6}$$

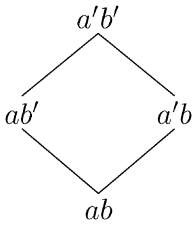
This type of rule is sometimes called *affirmation* in the literature (Bourne, 1966). The property a (= `blob_shaped`) in Fig. 1 is an example. Surely an important and salient part of that miniature world is the fact that all of its objects are blob-shaped—and there is a strong intuition that this will continue to hold in other objects generated from the same causal source. Indeed Medin, Wattenmaker, and Hampson (1987) found that subjects tend to sort objects in terms of simple, one-dimensional rules—that is, to extract "affirmations," whenever possible.

3.3. Algebraic combinations of regularities

Now consider the algebraic mechanics of our two concept classes, pairwise implication and constant properties. How do these concepts show up in the observable pattern of properties?

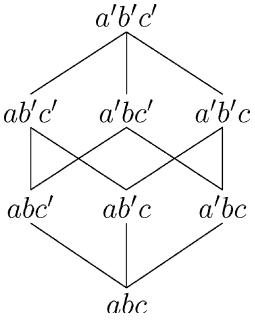
To answer this question, consider how a completely unconstrained set of observations is altered when each type of concept is introduced. A very informative way to examine the possibilities is to view them visually using a diagram. Because each possible object x can be regarded as a subset of Σ (i.e., the set of features that hold true in x), the set of possible objects is simply the set of all subsets of Σ. This can be illustrated in a diagram called a Boolean lattice (see Watanabe, 1969, 1985 for applications to concept learning). Such a diagram is constructed by placing objects with more properties true lower on the page than ones with fewer properties true, and connecting two objects by an edge whenever one is exactly the same as the other except for the truth value of one property. For two properties a and b, this lattice looks

like this:



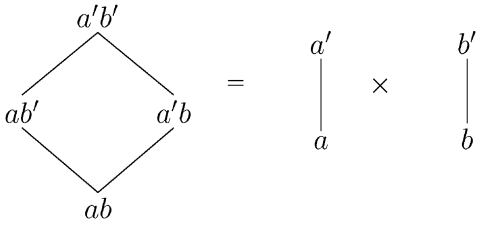
(7)

For three properties a, b, c, it looks like this:



(8)

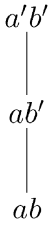
It is no coincidence that the above two diagrams look like a square and a cube, respectively; every completely unconstrained object set with D properties is isomorphic to the Boolean D-cube. Indeed, one can express the larger lattice literally as the product of smaller ones, e.g.



(9)

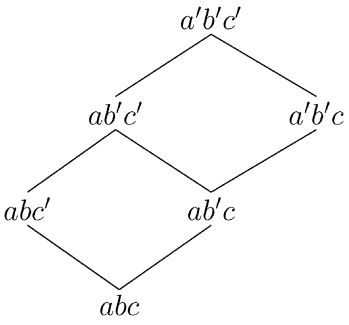
where the multiplication symbol $\times$ denotes the Cartesian product with an induced subset ordering (see Davey & Priestley, 1990 for an introduction). This equation expresses the fact that the unconstrained lattice for $\{a, b\}$ is the free (orthogonal) combination of the lattice for $\{a\}$ with the one for $\{b\}$.

What happens to these diagrams when some constraint is introduced? Imagine instead of all subsets of $\{a, b\}$, we have only those that obey a pairwise implication, say $b \to a$. The object $a'b$ is now prohibited, altering the structure of the lattice, which now becomes



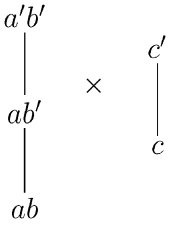
(10)

When the same constraint $b \to a$ is introduced on $\Sigma = \{a, b, c\}$ the lattice becomes



(11)

Inspection of this lattice reveals that it is actually equivalent to



(12)

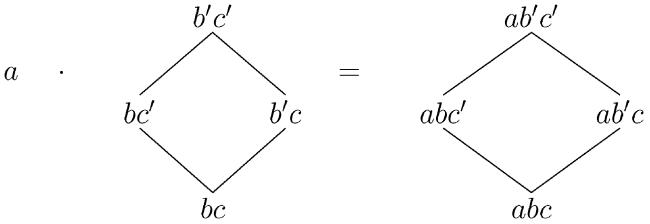
That is, the lattice for $\{a, b, c\}$ with $b \to a$ is the product of a constrained part ($\{a, b\}$ with $b \to a$) with a free part ($\{c\}$ with no constraint)—exactly as one would expect.

Object sets with pairwise implicational constraints have a very special kind of form; not just any combination of objects corresponds exactly to this kind of structure. (In fact, as D increases without limit, such combinations make up an ever smaller percentage of the total number of possible combinations, 2^{2^D}.) Technically, a lattice corresponding to a pure implicational constraint is *distributive*. (See Davey & Priestley, 1990 for an introduction to distributivity; or Erné, 1993 or Feldman, 1997 for discussion of the psychological significance of this property.)

What about the other kind of constraint, the constant property? Consider an object set $\mathscr{W}|_\Sigma = \mathbf{x} = \{x_1, x_2, \ldots\}$. When a single new property σ (not in Σ) is added (or newly recognized), it is simply concatenated (conjoined) with each member of $\mathbf{x}$:

$$\sigma\mathbf{x} = \sigma x_1 + \sigma x_2 + \cdots. \tag{13}$$

Clearly, this does not change the number of objects in $\mathbf{x}$, nor, critically, does it change the relationships among the various objects, as (for example)



(14)

(where the dot (·) indicates conjunction). The addition of a constant property does not change the structure of the lattice; it only changes the "spelling" of the nodes.

Summarizing, the two types of patterns introduced so far, affirmation and implication, serve as conceptual "atoms." Jointly, in algebraic combinations, they suffice to express arbitrary distributive featural patterns; which means that taken together this algebraic language can describe any set of objects whose features are entirely the product of quasi-causal implications (plus some constant component). In addition, there may also be an *a*causal component—the set of properties in Σ that are neither ever-present (affirmed) nor involved in any pairwise implications. These properties are freely combined with the structure that holds over the other properties, which means that the full lattice is the Cartesian product of a constrained part and an unconstrained (Boolean) lattice.

Hence the conceptual structure built up so far reduces to the following algebraic form. Write α for the set of constant properties; ω for the set of pairwise implications; and β for the rest of Σ, the unconstrained properties. Every possible world that can be produced by these types of concepts alone can be expressed as the Cartesian product of the lattice for ω with the lattice for β, conjoined with the properties α, that is

$$\alpha[\beta \times \omega]. \tag{15}$$

Now, working backwards from an observed world $\mathscr{W}$, projected into a property set Σ, the observer would seek to find the (unique) minimal solution α, β, ω to the expression

$$\mathscr{W}|_{\Sigma} \subseteq \alpha[\beta \times \omega]. \tag{16}$$

This solution represents a natural causal interpretation of the observed world $\mathscr{W}$ given the concept types proposed, including a component of *constant properties* α, a component of *causally irrelevant properties* β, and a component of *causally interacting properties* ω. It is really ω that captures the structure in the situation: an empty ω indicates an almost completely structure-free environment in which all objects are α's but otherwise all properties interact orthogonally.

The fact that Eq. (16) is an inequality rather than an exact equality reflects the fact that, as mentioned, not all sets of objects exactly correspond to any perfect causal (distributive) decomposition (see Example 1). In such a case, the algebraic solution to Eq. (16), the interpreted world, would exhibit a more coherent structure than the world itself—a "regularized" interpretation. Such an interpretation would reflect the observer's bias towards a more structured interpretation—a bias strong enough to override some degree of disagreement with the data itself.

It is instructive to consider the resemblance between Eq. (15) and the equation for a straight line (or plane or hyperplane). In this analogy, the term α plays the role of the constant intercept; the implication set ω plays the role of slope; and the orthogonal term β extrudes the line out into additional dimensions. In this sense, the interpretation of structure entailed by Eq. (15) can be thought of as the observer regressing a set of objects to a linear (i.e., pure-implication) model. It turns out that this metaphor carries through mathematically in an interesting way, and leads to a powerful generalization of the concept algebra that will be developed in the next section. Henceforth pure-implication concepts (i.e. concepts expressible in the simple algebra) will be referred to as *linear*.[3] The empirical hypothesis—discussed more fully below but useful to keep in mind during the following examples—is this:

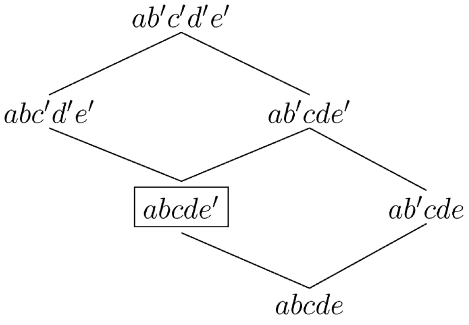
Linearity hypothesis. Human learners are biased towards linear concepts.

This hypothesis will be elaborated and supported by empirical evidence below. First, though, some simple examples are useful to get the general idea of how the algebra works.

Example 1 (*The amoeba world*). Consider again the amoeba world of Fig. 1a. The minimal solution here is

$$\begin{aligned} &\alpha = \{a\}, \\ &\beta = \{b\}, \\ &\omega = \{e \to c, c \to d, d \to c\}. \end{aligned} \tag{17}$$

This corresponds to the lattice:



(18)

which decomposes to

$$a \cdot \left[\begin{array}{ccc} & & c'd'e' \\ & & | \\ b' & & \\ | & \times & cde' \\ b & & | \\ & & cde \end{array} \right] \tag{19}$$

In either form, the solution has a natural interpretation. The α component means that all objects in this world are `blob_shaped` (a). The regularity that all objects in this

[3]The definition of *linearity* here (discussed more fully below) is not directly related to the notion of *linear separability* of categories, which refers to the separability of positive and negative examples by a hyperplane. Linearly separable categories have generally not been found to be easier for humans to learn (Medin & Schwanenflugel, 1981; Wattenmaker, Dewey, Murphy, & Medin, 1986); though for a contrary view see Minda, Smith, and Morgan, (1997).

world are drawn from the same general type (blob-shaped objects), which is intuitively obvious to the human observer, is captured by the algebraic solution in a transparent way. The β component means that being `shaded` has no causal meaning—this property is not a cue to any other structure in this world.

The ω component is most interesting. It contains a *cycle*, i.e. a set of properties that imply each other:

$$c \to d,$$
$$d \to c,$$

in other words,

$$\texttt{has\_nucleus} \to \texttt{has\_dotted\_membrane},$$
$$\texttt{has\_dotted\_membrane} \to \texttt{has\_nucleus}.$$

This cluster of mutually correlated properties suggests a "mode," or, one might say, a *species*: a subpopulation in which certain properties consistently co-occur (Quine, 1985; Rosch & Mervis, 1975). Secondarily, some though not all of these dotted-membraned, nucleated objects are `large`; again this is not an arbitrary relationship, but rather the former is a precondition of the latter.

Notice that according to the inferred interpretation, one node in the above lattice, indicated by a box in Eq. (18), is actually *missing* from the observations—reflecting the fact that the linear (distributive) model of the world in this case did not capture the observations perfectly. That missing category, the observer infers, is possible under the underlying causal structure of this world, and ought to occur occasionally. This illustrates how the expressive weakness of the algebra enables the observer to make strong inductive inferences about the world, and how these inferences are regularized with respect to observations.

Example 2 (*A "Bongard" problem*). A second example is provided by the classic visual induction problems of Bongard (1970), often held up as a benchmark of intelligent generalization. Each problem consists of 12 panels such as those shown in Fig. 2, six on the left and six on the right. The problem posed to the observer is to generalize from the examples given to the presumably infinite classes exemplified by the two groups, somehow ignoring the concrete but nevertheless uninteresting distinctions among panels on each side (e.g. some are larger, some smaller, some shaded, some not, etc.).

The dichotomy between the two groups is conveniently labeled (*left* and *right*) by an oracle, thus making this an example of a "supervised" induction problem. The essential problem here is to determine the correct properties suitable for distinguishing the two groups—that problem is not addressed here. Rather, we focus on how the algebra expresses the structure inherent in the Bongard set-up. The canonic Bongard problem is in essence constructed as follows: 12 objects of some general type a, of which the six on the left have some property x, while the six on the right have x'. On top of this, some distractor

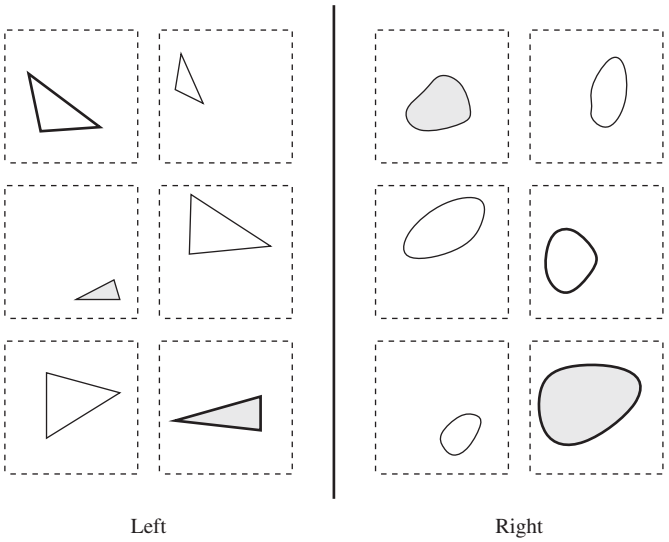


Fig. 2. A typical "Bongard problem" (after Bongard, 1970). What is the difference between the left and the right?.

properties b are added, which interact orthogonally with x on both left and right sides.

Of course this composition is reminiscent of the composition of the concept algebra itself, and it is no surprise that the algebra captures it neatly. To see how, consider that the oracle's labels *left* and *right* can be regarded as another feature, coded as l and l' (though note that learners may in reality treat labels differently from other features; see Love, 2002.) Now the Bongard problem can be regarded as a world $\mathscr{W}_B$ in exactly the sense defined above. The structure of this world is captured when it is projected into the alphabet $\Sigma_B = \{a, b, l, x\}$, in that

$$\omega(\mathscr{W}_B|_{\{a,b,l,x\}}) \supseteq \{l \to x, x \to l\}. \tag{20}$$

The cycle $l \leftrightarrow x$ contained in the solution captures the fact that in the Bongard world the critical property x consistently occurs on the left, and consistently fails to occur on the right. Moreover, this statement holds for any larger alphabet $\tilde{\Sigma}_B \supseteq \Sigma_B$. The structure inherent in the Bongard world is captured cleanly in ω-space.

Example 3 (*An unlabeled Bongard problem*). The above analysis of the labeled (supervised) Bongard problem reduces the oracle's labels *left* and *right* to a property in Σ. Hence it makes sense to regard an *un*labeled (unsupervised) problem isomorphically, just so long as there is at least one property playing the role of the label l in the labeled case (though again see Love, 2002). All that is required is that this property exhibit a consistent correlation (mutual implication) with the critical property x. The structure inherent in the world plays the role of oracle.

Consider the new population of amoeba-like objects shown in Fig. 3. After a bit of inspection, it is intuitively clear that there are two categories of object here: one whose members are consistently c and e (i.e. `has_nucleus` and `large`), and other whose members are consistently not.

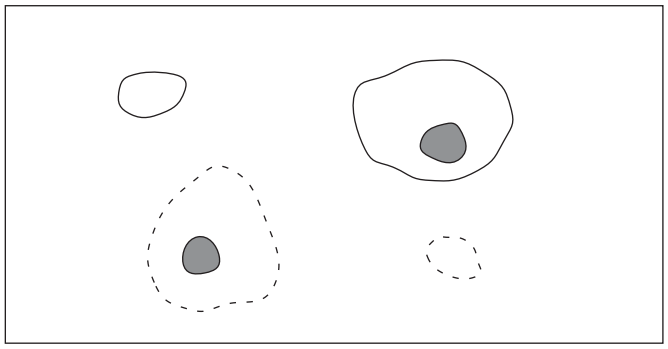


Fig. 3. An alternative population of "amoebae," amounting to an unlabeled Bongard problem.

Feature d (`has_dotted_membrane`) plays no role in this world and hence appears in β. Algebraically, the full solution is

$$\alpha = \{a, b'\},$$
$$\beta = \{d\},$$
$$\omega = \{c \to e, e \to c\}. \tag{21}$$

This is a fairly structured solution—i.e. it has a large ω. In particular it has a 2-property cycle that clearly corresponds to a mode or subspecies, and points to a systematic distinction of exactly the same type as the left–right distinction in the labeled Bongard problem. The featural structure inherent in the problem, rather than any overt labels provided by an oracle, provides the inferential leverage needed to recover the categories.

4. The generalized concept algebra

The previous section proposed an algebra that models simple featural patterns as, in effect, the sum of a constant plus a linear term. It is natural to ask: what is the analog of quadratic, cubic, or higher-order terms in this framework? What do we need to add to represent more complex patterns? This section gives an answer to this question, giving a complete hierarchy of terms of different degrees of complexity. This places the simple algebra in a more complete formal context, and makes its mathematical characteristics more clear. It also makes it possible to grapple with the very basic question of why (and whether) simpler concepts, such as linear ones, are more helpful than complex ones for the observer's goal of inferring the true structure of the world.

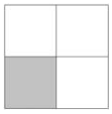
4.1. The power series expansion of a concept

The "atomic" concepts from which linear concepts are built were of two forms: constant (ever-present) properties, which involve one feature at a time; and pairwise implications, which involve two. This section develops a generalization to regularities that involve an arbitrary number of features at a time.

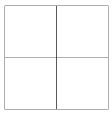
To motivate the generalization, consider again the structure of a pairwise implication:

$$a \to b. \tag{22}$$

The truth table for this concept looks like this:

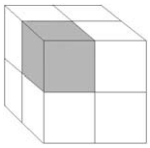


Here white space indicates a legal pair ("true") and shaded space denotes a forbidden pair ("false"). By contrast, the completely unconstrained relationship between two properties has a perfectly blank truth table:



Viewed this way, it is easy to see that pairwise constraint imposes the *minimal* amount of structure on the relationship between a and b—just one forbidden cell, namely ab'. This sense of minimal imposition of constraint is important if more complex concepts are to be built by putting together minimally structured patterns.

Hence a natural extension of pairwise implication to three variables is the $2 \times 2 \times 2$ truth table that looks like this:



corresponding to an implication of the form:

$$ab \to c. \tag{23}$$

As in $a \to b$, exactly one cell is shaded. Notice that the labels on this table are omitted. Any implication with a truth table of the same form—i.e. that can be made equivalent to the above simply by spinning it in space—is in a sense equivalent. It is very convenient to be able to discuss classes of concepts that are equivalent in this way, sometimes called *congruence*. Formally, two concepts are congruent if one can be transformed into the other simply by the interchange of property labels and property polarities (assignment of negative and positive values; see Feldman, 2000, 2003a; Shepard et al., 1961). Intuitively, two concepts are congruent if their truth tables are rigid rotations of each other. Writing $\psi_1 \cong \psi_2$ to indicate that the two concepts ψ_1 and ψ_2 are congruent, we have, for example,

$$(a \to b) \cong (a \to b'). \tag{24}$$

This example should make it clear that congruence is not the same as logical equivalence. All "constant" concepts, including any σ or σ', are congruent to the concept a,

because they have all the truth table:

Hence our generalized atomic concept with an arbitrary number of variables always has a truth table with exactly one shaded cell. To see what symbolic form such a concept would take, consider that a pairwise implication $a \to b$ can also be written as

$$a' + b \tag{25}$$

which is equivalent to

$$\neg(ab'), \tag{26}$$

(for clarity writing $\neg$ for negation outside the parentheses); which is congruent (though not equivalent!) to

$$\neg(ab). \tag{27}$$

Similarly, $ab \to c$ is congruent to

$$\neg(abc). \tag{28}$$

(This is the direct symbolic equivalent of the phrase "one shaded cell" in the three-feature case). The general case is an implication of the form

$$\sigma_1\sigma_2\cdots\sigma_K \to \sigma_0, \tag{29}$$

with K antecedent conjuncts and one consequent, congruent to the form

$$\neg(\sigma_1\sigma_2\cdots\sigma_{K+1}). \tag{30}$$

The number K here, called the *degree* of the concept, is the number of properties that are antecedent to the implication, which serves as a measure of the *complexity* of the regularity. For constant regularities $K = 0$, and for pairwise-implication regularities, $K = 1$. Concepts of this form (Eq. (29)) will be called *implication polynomials*, and a polynomial of degree K denoted by ϕ^K. Summarizing, we have

$$\begin{aligned}
\phi^0 &\cong \neg(\sigma)\\
\phi^1 &\cong \neg(\sigma_1\sigma_2)\\
&\vdots\\
\phi^K &\cong \neg(\sigma_1\sigma_2\cdots\sigma_{K+1}).
\end{aligned} \tag{31}$$

As discussed above, constant and linear regularities together do not express all Boolean concepts. But when concepts of arbitrary degree are added, all concepts can be expressed—in fact, *any set of observations can be expressed completely as the product (conjunction) of the implication polynomials that it satisfies.* This crucial fact—that every concept has a representation in the concept algebra—is captured by the "representation theorem" presented below.

It is natural to ask *how many* rules of each degree a given set of objects satisfies, but there is technical obstacle to answering this question given the definition above. The complete set of implication polynomials of all degrees satisfied by a given world may be highly redundant, simply because the set of rules that hold is, necessarily, transitively closed.[4] For example, if $a \to b$ and $b \to c$ hold then so does $a \to c$, and hence the latter is in the full concept set along with the former. Similarly, if a constant property a holds, then so do $b \to a$, $c \to a$, etc., simply by virtue of the fact that a entails $a + b'$. Hence every "real" rule that holds is accompanied by a host of automatically entailed additional rules, giving a false sense of how many rules the set of observations actually obeys. Instead, in order to get a more correct sense of how many *independent* rules are satisfied, we would like to pull out only those rules that are absolutely required to express the observations. Such a set, called *irredundant*, includes only the "real" rules the observation set obeys, omitting those that are obeyed only because some other rule or rules entails them.

The set of regularities that a given concept obeys, pruned of redundancies in this sense, will be called a *power series expansion* of the concept. The power series is (in a sense more fully described in the Appendix) the simplest way of describing the concept as an algebraic combination of implicational rules. Theorem 1 demonstrates that for every concept such as minimal representation exists.

The term "power series expansion" reflects the fact that the representation "unfolds" or expands the observation set into a layered collection of rules of different degrees (K) of complexity, in roughly the same way that the Fourier series expands a periodic function into a sum of sine and cosine components of various frequencies, or a Taylor series expands an analytic function into a combination of derivatives at various levels of degree. For an object set $\mathbf{x}$, the set of polynomials of degree K that are contained in the series expansion of $\mathbf{x}$ will be denoted by $\Phi_{\mathbf{x}}^K$. Of course for a particular set $\mathbf{x}$ and a particular K, $\Phi_{\mathbf{x}}^K$ may be the empty set—no regularities of degree K apply. The full power series for observations $\mathbf{x}$ (i.e., the union of $\Phi_{\mathbf{x}}^K$ for all K) will be denoted $\mathscr{S}(\mathbf{x})$.

The power series $\mathscr{S}(\mathbf{x})$ in full form looks like this:

$$\begin{aligned}
0 &: \Phi_{\mathbf{x}}^0 \quad \text{(minimum degree)}\\
1 &: \Phi_{\mathbf{x}}^1\\
&\vdots\\
K &: \Phi_{\mathbf{x}}^K\\
&\vdots\\
D-1 &: \Phi_{\mathbf{x}}^{D-1} \quad \text{(maximum degree)}
\end{aligned} \tag{32}$$

The representation theorem simply says that every object set $\mathbf{x}$ has a power series expansion $\mathscr{S}(\mathbf{x})$.

[4]Note that here the translation to fuzzy probabilistic rules is not completely straightforward, because the rule $p(\sigma_1|\sigma_2) > \theta$ is not transitively closed. For further discussion of the stochastic case see the section below on "Estimation of the power series from noisy data".

Theorem 1 (*Power series representation*). *For any object set* $\mathbf{x}$ *defined over a language* Σ *of size* D, *there exists a minimal irredundant set of polynomials* $\mathscr{S}(\mathbf{x})$ *such that*

$$\mathscr{W}|_{\Sigma} = \mathbf{x} = \Phi^0_{\mathbf{x}}\Phi^1_{\mathbf{x}}\cdots\Phi^{D-1}_{\mathbf{x}} = \bigwedge_{K=0}^{D-1}\Phi^K_{\mathbf{x}} = \bigwedge \mathscr{S}(\mathbf{x}), \tag{33}$$

where $\Phi^K_{\mathbf{x}}$ denotes the set of polynomials in $\mathscr{S}(\mathbf{x})$ having degree K. (A proof of this and other theorems is given in Appendix A.) (For a given concept, there may be multiple equivalent irredundant polynomial sets; among these the power series is one with the minimum power spectrum—see below.) The power series is thus a complete but irredundant expression of $\boldsymbol{x}$ in terms of the regularities contained within it, divided up into regularities of different levels of internal complexity.

Connecting this up with the "simple algebra" of the previous section, note that what was called $\alpha(\mathbf{x})$ (the ever-present features) is really just $\Phi^0_{\mathbf{x}}$ (the regularities of degree 0); and what was called $\omega(\mathbf{x})$ (the pairwise implications) is really just $\Phi^1_{\mathbf{x}}$ (the regularities of degree 0). $\beta(\mathbf{x})$, the part of Σ not mentioned in either $\Phi^0_{\mathbf{x}}$ or $\Phi^1_{\mathbf{x}}$, were previously indescribable, but might now be expressed by some combination of higher-order terms in the power series. While the simple algebra could only express linear concepts, the full power series can describe any discrete-featured pattern.

4.2. The power spectrum

As suggested, it is now possible to ask how many regularities of degree K a given concept contains, a number indicating how much of the concept's structure is of degree K. By analogy with other types of power series, this number is called the *power* of $\mathbf{x}$ at degree K, and denoted $\lambda^K_{\mathbf{x}}(=|\Phi^K_{\mathbf{x}}|)$ for brevity (where the notation $|S|$ denotes the number of elements in set S). For a particular $\mathbf{x}$ the function tabulating the power $\lambda^K_{\mathbf{x}}$ at each degree K is called the *power spectrum* of $\mathbf{x}$. Specifically, we use the (unique) minimal power spectrum among all irredundant regularity sets equivalent to the concept (see Appendix A). The power spectrum is simply a list of number pairs:

$$\begin{aligned} &0 : \lambda^0_{\mathbf{x}} \quad \text{(minimum degree)}\\ &1 : \lambda^1_{\mathbf{x}}\\ &\vdots\\ &K : \lambda^K_{\mathbf{x}}\\ &\vdots\\ &D-1 : \lambda^{D-1}_{\mathbf{x}} \quad \text{(maximum degree)} \end{aligned} \tag{34}$$

and thus can be conveniently plotted, giving a compact graphical summary of the degree of regularity of the observations—that is, now much of the featural variation can be accounted for by simple (low-degree) rules, and how much by more complex (higher-degree) rules. At one extreme, an observation set in which all objects have feature a, but obey no other patterns, would have all its power at degree 0. At the other extreme, an observation set with all its power at maximal degree $(D-1)$ is maximally complex in that it obeys no simpler rules at all, i.e. no rules of degree less than $D-1$. Most concepts will fall somewhere in between. Linear concepts have all their power at $K=0$ and $K=1$.

4.3. Algebraic complexity

Occasionally it will be useful to use a single scalar measure of spectral complexity, called the *algebraic complexity*, and defined as the total power weighted by a function w_K of degree λ^T:

$$\lambda^T = \sum_{K=0}^{D-1} w_K \lambda^K_{\mathbf{x}}, \tag{35}$$

where the weights w_K are linear increasing in K and sum to zero (i.e. defined by $w_K \propto K$, $\sum w_K = 0$, and $\sum |w_K| = 1$). This single number gives a simple measure of how much of a concept's power falls higher in the spectrum, and hence serves as a measure of the concept's complexity as represented in the algebra. As previously mentioned, it is known that learnability decreases with conceptual complexity; indeed this was the major conclusion originally drawn in Feldman (2000). Other recent research has shown similar complexity effects (Fass & Feldman, 2002; Pothos & Chater, 2001, 2002) with concepts defined over other types of features. But of all the complexity measures in the current literature, algebraic complexity is the only one directly applicable to concepts defined over n-valued discrete features.

In the formula for algebraic complexity, other choices of weight functions w_K are possible. One particularly useful alternative is to weight each polynomial ϕ^K by $K+1$ (degree plus one), so that the algebraic complexity is equal to the total number of literals (mentions of variable names) in the power series. (Recall that each polynomial of degree K contributes $K+1$ literals to the power series, because each polynomial has the form given in Eq. (29).) Regardless of the weighting scheme, the main idea is that the weight assigned to a polynomial increases with its degree, so that greater complexity is associated with higher-degree polynomials, which describe finer or more detailed aspects of the overall pattern.

Matlab code for computing the power series, power spectrum, and algebraic complexity of arbitrary discrete-featured concepts is publicly available at http://ruccs.rutgers.edu/~jacob/demos/algebra.html. See Appendix A for a brief description of the algorithm implemented there.

4.4. Examples

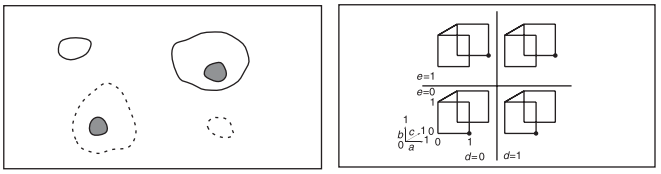
Figs. 4 and 5 show two different worlds along with their power series and power spectra. Fig. 4 shows the same unlabeled classification problem from Fig. 3, while Fig. 5 shows a highly random or nonlinear world. The easily solved classification problem has all its power at $K = 0$ and $K = 1$, while the random world has all its power at higher degrees. This correlation between spectral complexity and subjective complexity will be seen again in the empirical data discussed below.

4.5. Non-Boolean features

So far, only Boolean (binary) features have been considered. We now consider the more general case of n-valued discrete features, which the algebraic machinery extends naturally to handle.

Assume an alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_D\}$, in which each σ_i can now take on n_i distinct values $v_1 \ldots v_{n_i}$. That feature i takes on a certain value v will be denoted $\sigma_i(v)$, numbering the values of v from 0 to $n_i - 1$. For example, if feature a is *shape*, then we might have

$$\begin{aligned} &a(0) = \texttt{square},\\ &a(1) = \texttt{circle},\\ &a(2) = \texttt{triangle}, \end{aligned} \tag{36}$$



Φ^0: $\{a, b'\}$
Φ^1: $\{c \to e, e \to c\}$
Φ^2: $\{\}$
Φ^3: $\{\}$
Φ^4: $\{\}$

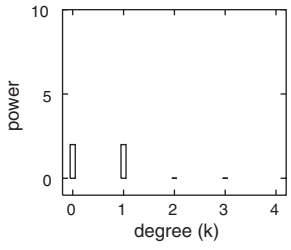
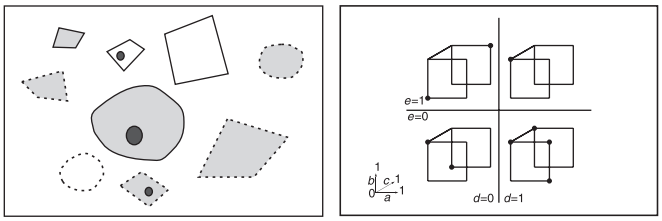


Fig. 4. The unlabeled classification problem from Fig. 3 (top, along with depiction in abstract Boolean 5-space), along with its full power series (middle) and power spectrum (bottom). All the power is at constant ($K = 0$) and linear ($K = 1$) degrees. Here $\mathbf{x} = \{10111, 10101, 10010, 10000\}$.



Φ^0: $\{\}$
Φ^1: $\{\}$
Φ^2: $\{c'e \to a', ce \to a', ce \to a, c'd' \to a'$
$cd \to a', b'd \to a, b'c \to a'\}$
Φ^3: $\{bd'e \to a, b'c'e' \to a, bcd' \to a\}$
Φ^4: $\{\}$

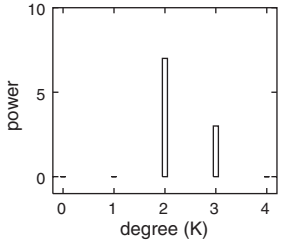
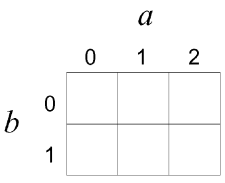


Fig. 5. A highly random or "nonlinear" world (top, with abstract illustration), full power series (middle), and power spectrum (bottom), showing how all the power is at higher degrees ($K \geqslant 2$). Here $\mathbf{x} = \{01011, 11101, 10010, 01010, 01110, 00001, 00100, 11010, 01000\}$.

with $n_a = 3$. Add to this a binary feature b, say *size* with values

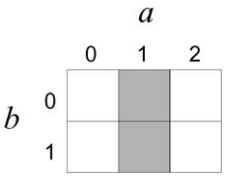
$$\begin{aligned} &b(0) = \texttt{large},\\ &b(1) = \texttt{small}, \end{aligned} \tag{37}$$

where $b(0)$ means what was previously denoted b' and $b(1)$ means b. This yields a 3×2 property language $\{a, b\}$, corresponding to the following truth table:



As before, each object corresponds to a cell in the table. Also as before, each implication polynomial corresponds to a pattern of shading in the table; the shape of the shading pattern depends on the degree K of the polynomial.

A "constant" ($K = 0$) concept, e.g. $\neg a(1)$, corresponds pictorially to one particular row or column of the truth table being shaded (prohibited), i.e.



or in English "no circles."

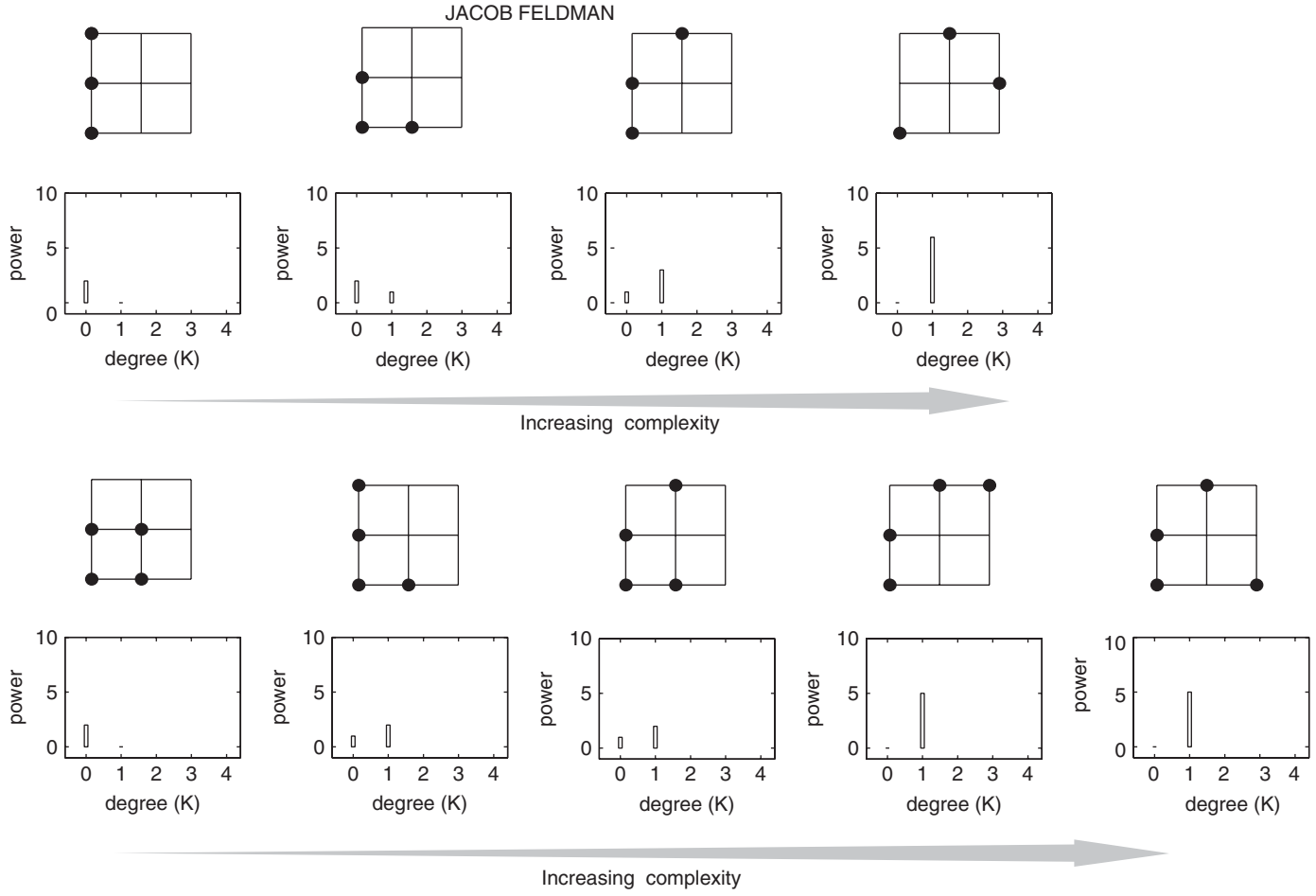
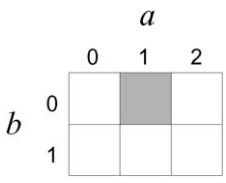


Fig. 6. Nine concepts in a 3×3 feature space along with their power spectra, illustrating the computation of spectra of concepts defined over features with more than two values each (cf. Lee & Navarro, 2002). The concepts in the upper row have three positives each, and those in the lower row have four each. Within each row, algebraically complexity increases from left to right (with a few ties), as can be appreciated by looking at the spectra.

A pairwise implication ($K = 1$), e.g. $a(1) \rightarrow \neg b(0)$, would look like



i.e. "all circles are small."

In general, an implication polynomial of degree K has the form

$$\sigma_1(v_1)\sigma_2(v_2)\cdots\sigma_K(v_K) \rightarrow \neg\sigma_0(v_0), \tag{38}$$

(analogous to Eq. (29)), with each v_i taking on a value drawn from the range of the variable σ_i.

With this definition, all the crucial properties of the algebra extend to the n-valued case. Specifically: all sets of objects have power series representations with unique minimal power spectra. The proof of this is a trivial extension of the proof in the Boolean case, hinging on the fact that every set of objects can be directly expressed as a conjunction of polynomials of maximal degree $D - 1$. As before, more regular worlds will have spectral power at lower degrees; only a completely chaotic and irregular world obeys only regularities of maximal degree. As illustration, Fig. 6 shows nine concepts defined over two three-valued feature along with their power spectra.[5]

4.6. Algebraic typicality of individual objects

In the algebraic treatment so far, an object either satisfies a set of regularities or utterly fails to, much as in traditional propositional calculus. However, the power series, because it divides the pattern up into distinct regularities, makes it possible to literally *count* the regularities that a given object on either side of a classification boundary satisfies. This opens the door to quantifying the degree of typicality exhibited by each object given a particular concept.

[5]This set of nine concepts, some of which were studied in the previously mentioned experiment of Lee and Navarro (2002), constitute an exhaustive classification of concepts in 3×3 space with three or four positive objects. Studies are currently underway in our laboratory measuring the learnability of these concepts.

Observe that among objects that do not obey *all* the regularities in a given world, some obey more than others. Some may observe *most* of the regularities in the given world, others none. This simple observation forms the basis of a route towards computing the "degree of fit" between a given object and the set of rules governing some world. In turn this forms the basis for predicting gradations of fit over a range of objects for a given category.

In general, for some power series $\mathcal{S}(\mathbf{x})$ derived from an object set $\mathbf{x}$, then for some object y, define the *typicality* of y, $\tau_{\mathbf{x}}(y)$, as the number of regularities in $\mathcal{S}(\mathbf{x})$ that y satisfies, i.e.

$$\tau_{\mathbf{x}}(y) = |\{\phi \in \mathcal{S}(\mathbf{x}) : y \text{ obeys } \phi\}|. \tag{39}$$

This number will be highest for y that are actually in $\mathbf{x}$, but for other objects will vary all the way down to zero (in the case of an object y that does not obey *any* of the regularities in $\mathcal{S}(\mathbf{x})$). In a very straightforward sense, the number τ indicates how much y fits into or "agrees with" the world described by $\mathcal{S}$, and thus serves as a measure of the degree of typicality of the object y taken as a member of the given concept. $\tau(y)$ forms a natural basis for predicting the degree to which objects will be judged typical of categories to which they may belong, and will be considered as such in the light of empirical data below.

4.7. Estimation of the power series from noisy data

So far the computation of the power series has been developed assuming perfect, noiseless data—i.e. the qualitative pattern among the features—as the emphasis has been on the composition of regularities and the structure of the power series. With real data, the observations $\mathbf{x}$ may reflect not only the regularity structure in $\mathcal{W}$, but also some stochastic process by which that structure is imperfectly realized in the data. Here we sketch a simple approach to estimating the power series from noisy data, loosely adapted from Fass and Feldman (2002). This scheme also allows the concept algebra to handle variations in the frequency of object instances over the various types, which are assumed to derive from stochastic sampling of the underlying regularity structure.

We first assume that $\mathcal{W}$ is governed by some "true" regularity set $\mathcal{S}$, which we wish to estimate. Unlike before, we assume that $\mathcal{S}$ is sampled stochastically (rather than deterministically via projection) to produce the observation set. Specifically, a series of N instances are generated from $\mathcal{S}$ in a sequence of N independent Bernoulli trials. Objects that *violate* $\mathcal{S}$ (i.e., fall in "forbidden" cells) are generated with some low probability ε, while objects that *obey* $\mathcal{S}$ are generated with probability $1-\varepsilon$; the parameter ε governs the magnitude of noise.[6] The resulting object set will contain high numbers of allowed objects, and low (but generally non-zero) numbers of forbidden objects, thus reflecting the regularity structure in a noisy manner. Recovery of the true generating power series $\mathcal{S}$ now reduces to a simple Bayesian estimation problem involving the prior probability $p(\mathcal{S})$ of each candidate power series and the likelihood $p(\mathbf{x}|\mathcal{S})$ of the observations $\mathbf{x}$ given $\mathcal{S}$.

There are $V = \prod_i^D n_i$ object types x, i.e. combinations of feature values (e.g. $V = 8$ in the Boolean 3-cube, and $V = 9$ in the 3×3 grid of Fig. 6). Assume that these types are canvassed with uniform probability $1/V$, with N objects in the resulting set, of which we count $|x|$ of type x. In effect we are making N throws of a V-sided die, with the "faces" (object types) having probability either $(1-\varepsilon)/V$ (in allowed cells) or ε/V (in forbidden cells). The likelihood of the observation set $p(\mathbf{x}|\mathcal{S})$ is determined by the probability of obtaining the observed counts $|x_i|$ under the candidate power series $\mathcal{S}$ (which determines the probabilities of the various types). Under these assumptions the ensemble of counts $|x_i|$ will follow a multinomial distribution (the generalization of the binomial distribution to more than two cases; see Box & Tiao, 1973),

$$|x_i| \sim \frac{N!}{|x_1|! \ldots |x_V|!} \left(\frac{1-\varepsilon}{V}\right)^P \left(\frac{\varepsilon}{V}\right)^{N-P}, \tag{40}$$

where P is the total count over all allowed cells. The expected number of objects in each allowed cell will be $(1-\varepsilon)N/V$, and in each forbidden cell $\varepsilon N/V$, with variance $\varepsilon(1-\varepsilon)N/V^2$ in either case. This distribution determines the likelihood $p(\mathbf{x}|\mathcal{S})$ of the observations under the candidate power series.

For the prior $p(\mathcal{S})$, following standard arguments in information theory, we assume that probability decreases exponentially with algebraic complexity λ^T; that is,

$$p(\mathcal{S}) \propto 2^{-m\lambda^T}, \tag{41}$$

in which the coefficient m controls the rate of decay of probability with complexity. This assumption penalizes (assigns low probability to) complex regularity structures with high spectral power, and favors (assigns high probability to) simple regularity structures with low spectral power.

Given these assumptions, the estimated power series $\hat{\mathcal{S}}$ will be the one that maximizes the posterior $p(\mathcal{S}|\mathbf{x})$, which by Bayes' rule is the one that maximizes the product of the prior and likelihood,

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{\operatorname{argmax}}\, p(\mathcal{S})p(\mathbf{x}|\mathcal{S}), \tag{42}$$

or alternatively minimizes its negative logarithm or "description length" (DL)

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{\operatorname{argmin}} -\log[p(\mathcal{S})] - \log[p(\mathbf{x}|\mathcal{S})] \tag{43}$$

(see Li & Vitányi, 1997; Rissanen, 1989). Plugging in Eq. (41), this reduces to

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{\operatorname{argmin}}\, m\lambda^T + \mathrm{DL}(\mathbf{x}|\mathcal{S}), \tag{44}$$

[6] An alternative assumption, perhaps more apt but also more complicated, would be to assign probability to each type x in proportion to its algebraic typicality $\tau(x)$.

where m is the decay coefficient used to define the prior (Eq. (41)). Thus, in sum, the estimated power series is the one that minimizes the weighted sum of the algebraic complexity plus the negative log likelihood of the observed counts under the multinomial loss function. In essence, this is the regularity set which is at once most plausible (simple), while fitting the observations as closely as possible given the assumed stochastic model.

5. Discussion

5.1. Interpretation of the power series

The power series of a given set of examples is a summary of the causal/implicational regularities, at all levels of degree, immanent in the data. This includes both broad, sweeping regularities, which have low degree, and relatively specific, fine, or "local" details, which have high degree, as well as everything in between. If the former predominate, the overall complexity will be low; if the latter, complexity will be high, meaning that the examples are relatively incompressible, and that a relatively large set of detailed statements are required to render it faithfully.

Thus the elements of the power series represent a kind of "theory" of the observed pattern—an instantaneous estimate, based on the pattern of observations, of which features influence which. It is known, however, that apprehension of such trends is influenced not only by observations, but also by prior knowledge and expectations (Heit, 2001; Pazzani, 1991; Wisniewski & Medin, 1994). Because each element of the power series represents a separate "belief" about a tendency in the world, one can easily imagine augmenting the power series by elevating or depressing its components in accordance with prior knowledge or expectations, perhaps in a Bayesian manner. Thus the computational machinery proposed above could be augmented by a more elaborate regularity-estimation mechanism that incorporated both observations and priors. Indeed this idea is very much in the spirit of the "theory–theory" motivation of the algebraic approach, and would represent a natural extension of the approach.

5.2. Connection to Bayes nets

This aspect of the power series also highlights another important connection: the similarity of algebraic power series to *Bayesian belief networks* or *Bayes nets* (Glymour, 2002; Pearl, 1986). Bayes nets are a formalism in which a potentially complex probability distribution is approximated in terms of the pairwise conditional dependencies it includes, often depicted in a diagram. Under certain conditions these dependencies have an intriguing interpretation as *causal* influences (Pearl, 2000). There has been a surge of interest in the statistical learning community in Bayes nets for a variety of applications, and they have recently been used to model psychological categories (Rehder, 1999; Rehder & Hastie, 2004). The formal

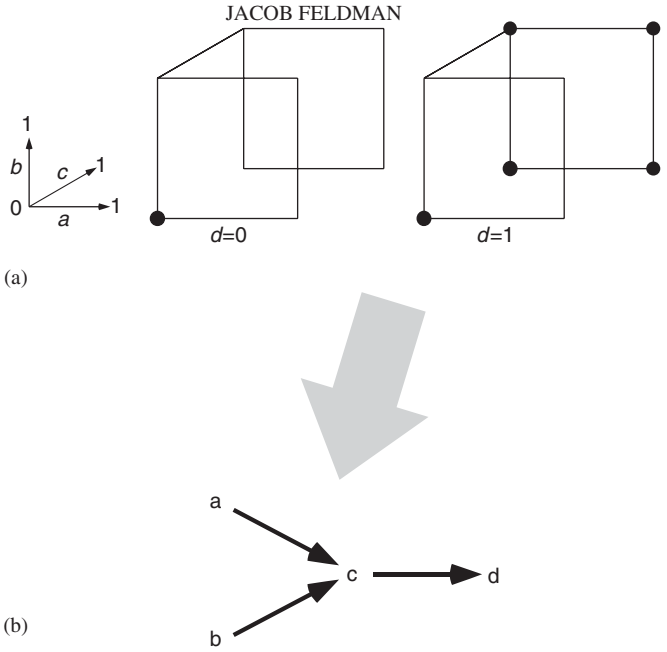


Fig. 7. (a) A four-dimensional concept and (b) its power series, depicted as a diagram analogous to a Bayesian belief network, illustrating the formal analogy between power series and Bayes nets. Here the power series consists of three degree-1 implications, $a \to c, b \to c$, and $c \to d$.

isomorphism between Bayes nets and power series is fairly straightforward. Essentially, the former decomposes a probability distribution into a product of distinct stochastic dependencies, while the latter decomposes a discrete featural pattern into a product (conjunction) of logical dependencies—also partialing them out into distinct levels of internal complexity (degree). Power series and Bayes nets are thus in a sense natural counterparts, one interpreting $a \to b$ as a probabilistic conditional dependency, the other as an instantaneous logical implication. In both cases the effect is to decompose the observations into a "theory"-like representation of the causal-like laws at work in the world from which the examples were generated (Fig. 7).

5.3. Successive approximation

As suggested above, the power series taken as a whole is not a good basis for induction, because it can represent any set of examples exactly, and thus represents everything equally well. (Technically, one would say it has no bias and is prone to overfitting.) But the separation of regularities into distinct levels of complexity (degree) opens up the possibility of representing patterns using a *truncated* series that contains only lower order or more basic regularities. Like a conventional power series expansion of a numeric function, the implication power series represents the set of observations as the combination (here, conjunction) of constant, linear, and higher-order terms. In both cases the expansion approximates the function with increasing

precision as more terms are added. When all terms are present (an infinite number in the Taylor series, D here) it represents the function exactly; but omitting higher-order terms yields successively more rough and qualitative approximations to the complete function. This division of the trends in the data to more and less essential elements resonates with evidence that human learners represent induced concepts as a "discrete structure plus noise," i.e. a simple rule-based core plus a more detailed and idiosyncratic periphery (Ahn & Medin, 1992; Medin, Altom, & Murphy, 1984; Nosofsky et al., 1994). Specifically the linear component of the power series is very similar to the network of dependency relations proposed by Sloman et al. (1998). In this spirit a very natural application of the power series is to vary the threshold of degree at which one approximates the observations (i.e. truncates the series) depending on circumstances (cf. the discussion of SUSTAIN below). A low threshold might be appropriate when one seeks only general trends, while a higher one might be required when a more detailed representation of the data is sought, perhaps after more extensive training. At low threshold only general trends can be appreciated, while including terms of all degrees allows one to in effect memorize the data—albeit with a concomitant cost in representational complexity and loss of predictive power.

5.4. Connection to hybrid models

Recently, many authors have proposed *hybrid* models that involve both rule-extraction (e.g., prototype formation) and exemplar storage (Ahn & Medin, 1992; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 2002; Medin et al., 1984; Nosofsky et al., 1994; Smith & Sloman, 1994). The common premise of these models is that human learners seek to extract common tendencies from examples, but also store individual cases, particularly exceptional ones. Such proposals implicitly involve an assumption that the rules involved are "simple," because there is no meaningful distinction between a complex rule and a heterogeneous collection of examples; any set of examples, no matter how disjoint, is formally equivalent to a single rule—albeit possibly a very complex one. Hence the rule-formation component of every hybrid model necessarily imposes some kind of constraint on the complexity of the rule. So the algebraic mechanisms developed in this paper may simply correspond to one half of a dual system—namely, the simple-rule-extraction part.

However, the generalization of the algebra to arbitrary levels of degree K means that the algebraic approach encompasses, in a sense, rules of all levels of complexity: from very simple constant or linear ones all the way up to degree $D-1$ rules that can describe any observation set, no matter how heterogeneous. Each regularity of maximal degree refers to one specific object, because it sets the values of all D features. Power series that contain only rules of maximal degree are, in effect, simply lists of individual objects, and in this sense are isomorphic to sets of exemplars. In this sense, representations in the concept algebra can range from extremely coherent and "prototype-like," at the low end of spectral power, to extremely heterogeneous and "exemplar-like," at the high end, with a full range of mixed possibilities in between.

Viewed this way, the concept algebra bears a striking affinity with a particular hybrid model, the SUSTAIN model of Love, Medin, and Gureckis (2004). SUSTAIN is a network model of human categorization in which the network grows new clusters whenever new observations cannot be fit into the existing ones. It is a hybrid model, in that it attempts at first to form low-dimensional rules (clusters), but will also store specific examples, by recruiting new clusters when new objects fail to fit into existing ones. Importantly, SUSTAIN is biased towards clusters of low dimensionality; it always attempt to fit new objects into clusters involving fewer features, only considering higher-dimensional clusters when simpler ones fail. The result after learning is a set of clusters of possibly varying dimensionality.

Thus SUSTAIN embodies both principles of minimality central to the concept algebra. Its eventual solution involves (a) the minimal number of clusters (in the algebra, implication polynomials), that (b) have the lowest possible dimensionality (in the algebra, degree). The algebraic power series shares precisely these properties (see proof of Theorem 1, Appendix A, for details). In this sense, both are hybrid theories, encompassing rules ranging all the way from the simple to the highly complex, with the former extreme approximating a prototype formation strategy, and the latter extreme approximating an exemplar storage strategy. Both attempt to find the minimum-dimensional rule combination that accounts for the observed objects, and resort to exemplar storage only to the extent that lower-dimensional (more prototype-based) strategies fail to account for observations.

Hence though superficially the two theories take very different forms, the concept algebra can be seen as a kind of analytic counterpart to SUSTAIN. The concept algebra gives an analytic understanding of and motivation for the process by which examples are broken down into rules and sub-rules of varying dimensionality, which otherwise might seem ad hoc. Conversely, as befitting a performance theory, SUSTAIN can account for some aspects of human data that the concept algebra cannot, in particular those aspects that stem from performance limitations.

6. Empirical evidence

There are a number of ways in which the theoretical ideas developed above might relate to human concept learning, of which we consider primarily two: (1) that algebraic complexity (Eq. (35)) corresponds to phenomenal or subjective conceptual complexity and difficulty in learning; and that (2) typicality in the concept algebra (Eq. (39)) predicts the judged typicality of objects within

categories. Along the way evidence that human learners have a special bias in favor of linear concepts (the linearity hypothesis) will be considered. No new data are presented in this paper; we rely primarily on the dataset of Feldman (2000), described below.

6.1. Subjective conceptual difficulty

6.1.1. Classical difficulty orderings

First of all, it is worth remarking that the well-known superiority of conjunctive over disjunctive concepts, as well as the other principal findings from the 1960s concerning the bivariate ($D = 2$) case (e.g. see Bourne, 1970; Haygood & Bourne, 1965), are all consistent with an analysis in terms of spectral content. Conjunctive polynomials are of minimal degree ($K = 0$), while disjunctive polynomials are always of maximal degree ($K = D - 1$). For example the conjunctive concept ab (in traditional notation, $a \wedge b$) if of degree $K = 0$, while the disjunction $a + b$ (i.e. $a \vee b$) is congruent to the polynomial $a' \rightarrow b$ of degree $K = 1$; and the disjunction $a + b + c$ is equivalent to the rule

$$a'b' \rightarrow c, \tag{45}$$

which is of degree $K = 2$.

Of course for $D = 2$, all concepts, including disjunctive ones, are linear. But disjunctive concepts still have higher algebraic complexity than conjunctive ones and are thus more phenomenally random. Biconditionals ($a \leftrightarrow b$) have even higher power, and indeed they are even more difficult for subjects to learn (Haygood & Bourne, 1965; Neisser & Weene, 1962). This essentially exhausts the set of bivariate concepts; for more extensive corroboration higher dimension must be considered.

In their study discussed above, Shepard et al. (1961) introduced a now-classic stimulus set that has often served as a basic test case for theories of conceptual difficulty. Shepard et al. considered objects defined over three features ($D = 3$), for a total of 8 ($= 2^3$) objects, of which half (four) were designated as positive examples and half negative. In this case there are exactly six distinguishable (i.e., non-congruent) types of concept, denoted by Shepard et al. as types I–VI. Illustrations of the six types can be found in the row marked 3[4] in Fig. 8, where they are labeled $\mathbf{1}_{3[4]}$–$\mathbf{6}_{3[4]}$. Shepard et al. found that the types ranked in subjective difficulty in the order I < II < [III, IV, V] < VI, with types III, IV, and V of approximately equal difficulty.

Among these six cases, the most regular, type I, is unique in having all its power at degree 0; the most irregular, type VI, is unique in having all its power at maximal degree $K = 2(= D - 1)$, and is in this sense *completely* unstructured. More generally, algebraic complexity correlates with the classic difficulty ordering ($r = 0.8433$). A slightly finer demonstration is to compare algebraic complexity against human data collected by the author on these concepts (as well as 70 others; see fuller description below). Correlation between complexity and proportion correct for the six concepts is $r = 0.7953$, slightly better than ALCOVE ($r = 0.7663$).

6.1.2. The dataset of Feldman (2000)

This agreement with the Shepard et al. (1961) finding is suggestive, but with only six types to compare is hardly probative, and at any rate this finding has been well-accounted for by several theories in the literature. But notwithstanding its fame, this set of six types represents only a small fraction of a much larger universe of concept types with different numbers of features and constituent members.

Data on a similar but much larger and more comprehensive concept set have previously been presented by the author (Feldman, 2000, previously mentioned in connection with the Boolean complexity effect). (See the earlier paper for a more detailed presentation of these data.) This larger concept set generalizes the Shepard et al. classification by using $D = 3$ or 4 and 2, 3, or 4 examples. This gives rise to a classification into 41 types, including Shepard et al.'s original six (see Fig. 8 for illustrations all 41 types, and see Feldman (2003a) for further extensions and discussion of this typology). It must be noted that this dataset does not precisely replicate the classic difficulty ordering over the famous six (see Fig. 10), suggesting some sensitivity to slight differences in methodology (e.g. examples presented in a group rather than one at a time). Indeed some variations in the classical difficulty ordering have occasionally been reported in the literature (e.g. Medin & Schwanenflugel, 1981). These cautionary notes concerning the Feldman (2000) dataset should be kept in mind below, but it should be understood that the empirical case for the algebra does not rest in any substantial way on its fit to the six Shepard types, but rather on the larger set of which they are only a subset.

Of the 41 cases, 35 (all but Shepard et al.'s original six) have distinct numbers of positives and negatives, and hence can appear in two versions: with the smaller set denoted positive (referred to as Up parity) or with the larger set denoted positive (Down parity). Up and Down versions were tested separately, yielding a total of 76 distinct concepts on which data were collected (i.e. 2×35 plus the six original Shepard et al. types, which have only one parity). This concept set thus represents a very wide range of discrete featural patterns, not only far more comprehensive than any other set previously tested, but in fact exhaustive: it includes *all* logically possible Boolean concepts with 3 or 4 features and 2, 3, or 4 positive examples. Thus this concept set avoids the possibility of bias in favor of any particular type of theory, as Smith and Minda (2000) argue has occurred within the literature due to over-reliance on certain concept sets.

In the experiments, stimulus objects were "amoeba"-like pictures defined by 3 or 4 salient binary features. In each concept block, a concept of the designated type was generated at random, first in abstract coordinates and then realized in terms of a randomly permuted assignment of

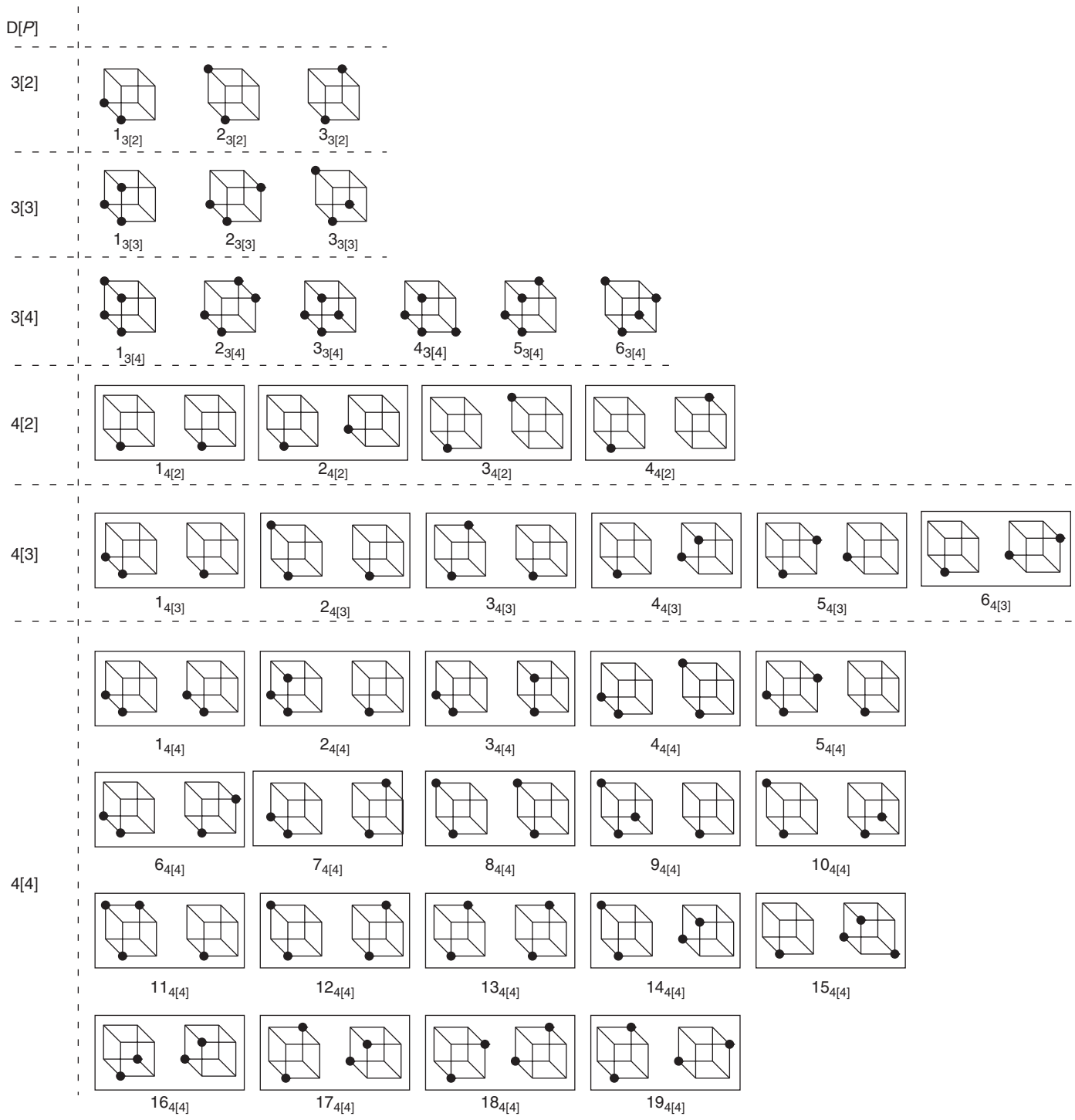


Fig. 8. Illustrations of the concept types tested in Feldman (2000), which comprise all possible logically distinct Boolean concepts with 3 or 4 features and between 2 and 4 positive examples (see Feldman, 2003a for discussion). Each figure shows Boolean 3- or 4-space with heavy dots on the vertices corresponding to positive examples. Families are labeled using the notation $D[P]$, indicating the set of concepts with D features and P positive examples. Individual concept types are labeled in arbitrary order subscripted with the family name, e.g. $\mathbf{1}_{3[4]}$, $\mathbf{2}_{3[4]}$ to indicate the first two cases in the family having 3 features and 4 positive examples. The row marked 3[4] is the family studied in Shepard et al. (1961). In Feldman (2000), each concept was tested in the version shown (called Up parity) and in a complementary version (Down parity), except for those in 3[4], whose complements are identical to the originals. This yields $76(= 6 + 2 \times 35)$ mathematically distinct concept types on which data were collected. Fig. 10 shows human performance for all the concept types displayed here, plotted with corresponding labels.

coordinates to actual features. During the learning phase on each concept, the subject was then shown all of the positive and negative examples at the same time, positive in the upper half of the screen and negative in the lower half (labeled as such), for a fixed duration ($2P$ seconds, where P is the number of objects in the smaller class, whether

positive or negative). Then, during the test phase, all 2^D objects were presented one at a time in random order, with the subject asked to classify them as positive or negative. Proportion correct then served as the main dependent measure, tabulated either over an entire concept, or for each individual object taken as a member of the concept in which it was presented. Frequency of presentation of individual objects was never varied in this dataset, as each object in each concept was simply presented once during the test phase of that concept. Hence this dataset makes an appropriate testing ground for the algebraic theory, which as discussed so far considers each concept in terms of the qualitative pattern of featural variation in contains.

First of all, as a relatively coarse initial comparison, linear concepts were learned better than nonlinear concepts (proportion correct = 0.876 vs. 0.795, $t(74) = 4.606$, $p = 0.000017$), corroborating the linearity hypothesis (Fig. 9).

A more complete test is to regress algebraic complexity λ^T (Eq. (35)) onto proportion correct (Fig. 10a). This regression is highly significant ($R^2 = 0.5037, F(1,74) = 75.11, p < 0.000001$). Patterns that have higher algebraic complexity are harder for subjects to learn, and the single number λ^T, without *any* free parameters or "fudge factors," accounts for more than half the variance in measured learning. (There are two degrees of freedom, slope and intercept, in the linear fit, but no free parameters at all in the computed complexity values.) Human observers are sensitive to the type of regular structure expressed by the concept algebra, and a given concept's difficulty in the mind of a human learner directly correlates with its degree of expressive complexity in the algebra.

6.1.3. Factors influencing conceptual difficulty

A more detailed examination of the Feldman (2000) data is to use the concepts' raw spectral components λ^0, λ^1, λ^2, and λ^3 as independent measures in the regression, rather than the single number λ^T ($= \sum_K w_K \lambda^K$). (With $D = 3$, λ^3 does not exist and is treated as zero.) This regression is also highly significant ($F(4,71) = 21.55, p < 0.000001$), and accounts for more than half the variance ($R^2 = 0.5484$).

A telling pattern is revealed when one looks at the standard coefficients of the four factors in this regression, i.e. the fitted weights in the regression of the contributions of the four λ^K's. The weights are 0.289, 0.011, −0.348, and −0.553, respectively, for $K = 0, 1, 2$, and 3 (plotted in Fig. 11). For $K = 0$ and $K = 1$, the coefficients are positive: the fuller and richer the linear component of the concept is, the better subjects can remember it. But for $K = 2$ and $K = 3$ the coefficients are negative: the more power a concept has at these degrees the *worse* subjects can remember it. More precisely, the weights drop steadily as K increases, crossing zero into negative territory above $K = 1$. This pattern accords precisely with the linearity hypothesis: human subjects tend to be helped or hindered by spectral power depending on whether it falls in the linear or nonlinear bands. Moreover, the almost perfect linearity in these weights retrospectively justifies the definition of λ^T, in which spectral components' weights are linear in K. Hence the mathematically natural definition of algebraic complexity is actually validated by the experimental data.

Another argument in favor of the algebraic theory emerges when one looks more closely at the factors controlling subjective difficulty in the Feldman (2000) dataset. In that dataset, the two orthogonal factors of Boolean complexity and parity combined to account for predict about half the variance in measured conceptual difficulties, about as well as the four spectral components discussed above. However, as discussed in Feldman (2000), neither the Boolean complexity effect nor the parity effect has any independent rationale or connection to a well-motivated inference theory. The complexity effect is, of course, broadly consistent with the long tradition of parsimony principles and simplicity biases in induction (Sober, 1975). But as discussed above, the propositional language underlying Boolean complexity has little psychological plausibility, and the specific compression scheme involved in estimating Boolean complexity has even less. By contrast, algebraic complexity is based on the combinations of implicational regularities, i.e., quasi-causal laws, and in this sense its motivation derives from fundamental goals of inference—to extract the meaningful regularities governing the observed objects.

Similarly, the parity effect is reasonable in and of itself, and indeed had been observed decades previously (Haygood & Devine, 1967; Hovland & Weiss, 1953; Nahinsky & Slaymaker, 1970). But it represents an additional assumption, because the Up parity preference does not in any way derive from a simplicity bias (again, complexity and parity are orthogonal); so it would be more parsimonious to do without it if possible. Exemplar models such as ALCOVE are also symmetric with respect to parity, and thus likewise cannot explain the parity effect.

However, the concept algebra explains both effects with no additional assumptions. Algebraic complexity and

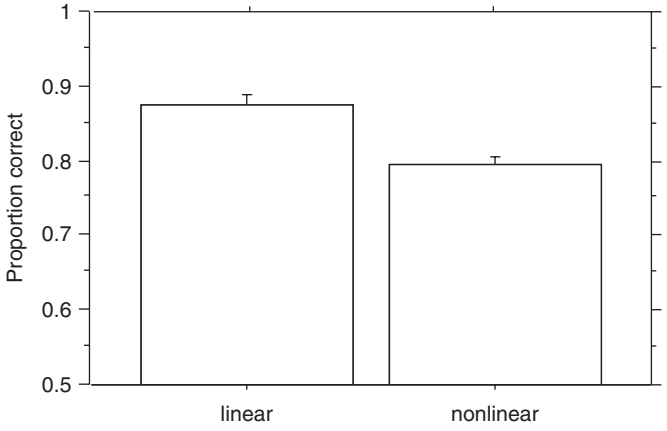


Fig. 9. Proportion correct for linear (left) and nonlinear (right) concepts.

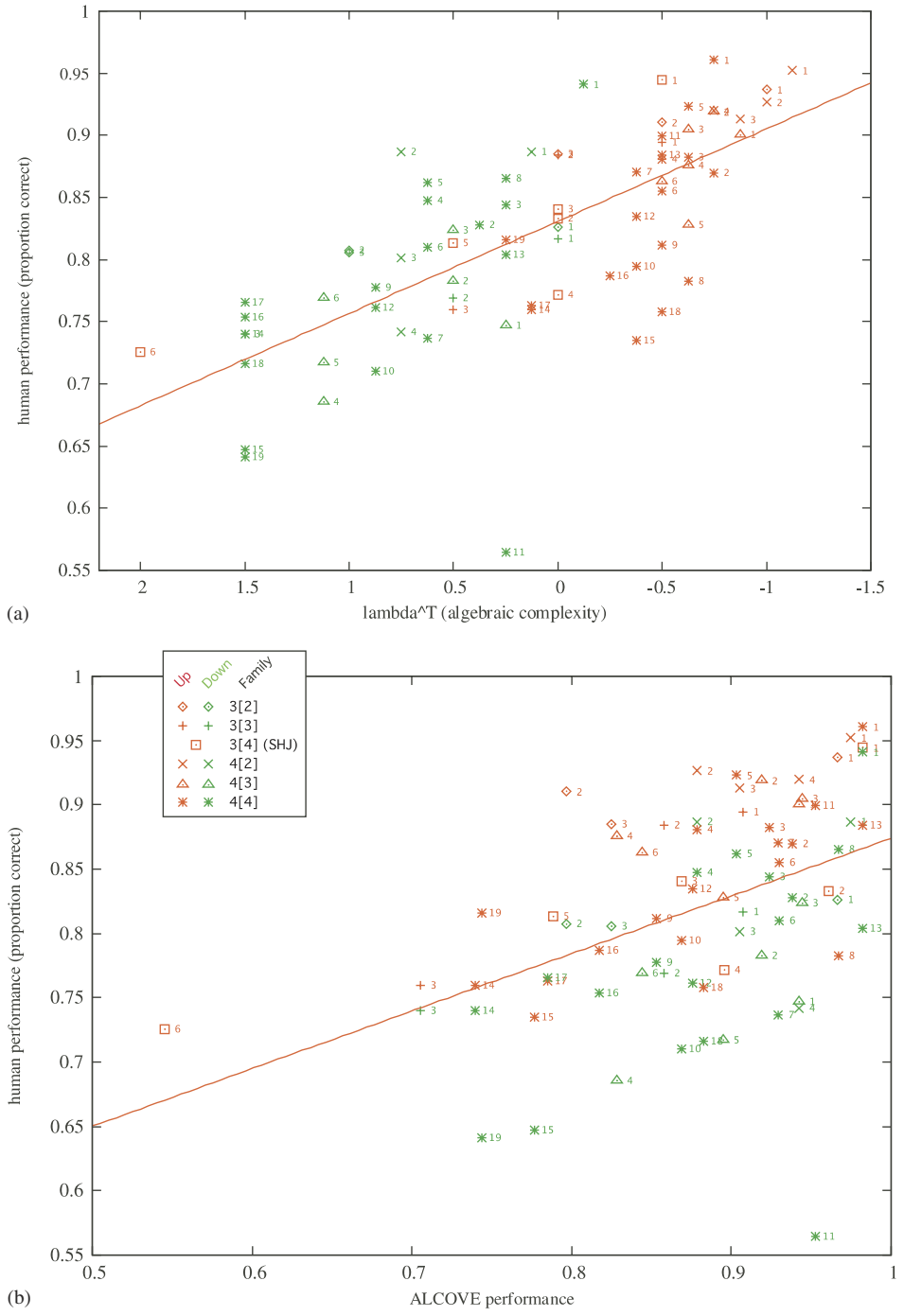


Fig. 10. Proportion correct for 76 concept types from Feldman (2000) plotted against (a) mean spectral power (algebraic complexity λ^T) and (b) performance of ALCOVE (Kruschke, 1992). Solid lines indicate linear regressions: for algebraic complexity, $R^2 = 0.5037$; for ALCOVE, $R^2 = 0.2119$. Note that the abscissa in panel (a) has been inverted to facilitate visual comparison with panel (b). Color coding indicates parity (up = red[dark], down = green[light]), symbol shapes indicate concept family, and the numbers on each symbol indicate the class label within each family. See Fig. 8 for illustrations of the concepts with corresponding labels.

Boolean complexity are correlated ($r = 0.43$ over the range of the 76 concept types), reflecting the fact that they are both measures of structural complexity; hence the two measures account for some of the same variance in the human data. Second, and more subtly, Down concepts tend to have more spectral power at higher degrees than Up concepts. That is, in a given random concept, more algebraic regularities tend to occur by accident in the smaller half (labeled positive in Up concepts, negative in Down concepts) than in the larger half (where the labeling is the reverse). Hence it is fair to conclude that neither the Boolean complexity nor parity effects are primary; rather,

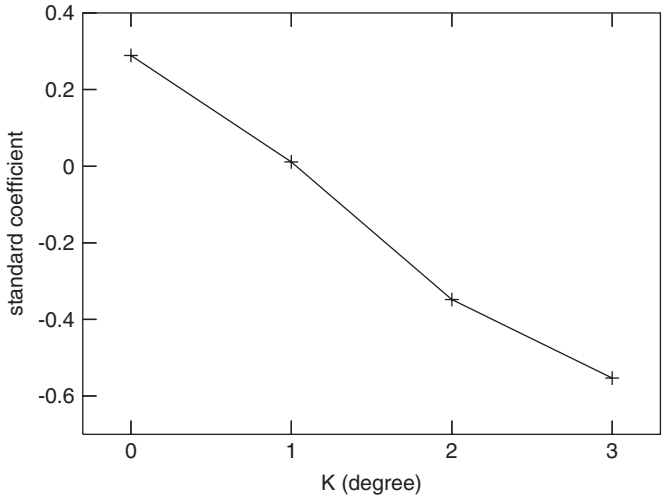


Fig. 11. Standard coefficients in the regression as a function of degree K, showing the linear decrease in contribution with increasing degree. The contribution crosses into negative territory when degree becomes non-linear.

both emerge empirically as side-effects of spectral power. And unlike these two effects, spectral power has an independent motivation, namely as a measurement of the regularity content of the data.

6.1.4. A comparison with ALCOVE

Given the heavy reliance of the concept algebra on the extraction of regularities from the data, it is particularly essential to compare its predictions to a model that involves no regularity-extraction, such as an exemplar model. The ALCOVE model of Kruschke (1992) is a very successful and representative exemplar model, a network implementation that has been found to fit human data particularly well (e.g. Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994). Philosophically, the concept algebra and ALCOVE make a sharp contrast. In the algebraic theory, the main thrust is the discovery of regular trends in the observations, while in ALCOVE (as in other exemplar models), there is *no* overt abstraction or extraction of regularity per se from the data. The two theories differ in several more concrete senses, too. To its credit, ALCOVE can handle continuous features (though, as mentioned above, Lee & Navarro (2002) found that it could not fit discrete-featured concepts well without being customized with a feature-based metric). Conversely, ALCOVE has a somewhat high number of parameters controlling its behavior, while the concept algebra in pure form has no intrinsic free parameters at all. (As mentioned, when algebraic typicality is fit to data, there are normally two extrinsic parameters, the slope and intercept of the linear fit, but the theory itself has no parameters.)

As discussed above, Smith and Minda (2000) have argued that the empirical success of ALCOVE and other exemplar theories may actually stem in part from historical over-reliance on certain small and unstructured concepts that may have favored exemplar models. The Feldman (2000) concept set is expressly designed to avoid such bias, in that it includes *every* logically possible concept over a certain well-defined range. Hence it represents an objectively level playing field in which to compare theories, including all concepts with all possible degrees of inherent structure.

Fit of ALCOVE to the Feldman (2000) data is shown in Fig. 10b. The figure shows the human performance (proportion correct) plotted against ALCOVE's performance (proportion correct) after 500 epochs (for $D = 3$ cases) or 1000 epochs (for $D = 4$).[7] In this computation, ALCOVE's various parameters were set as indicated by Kruschke (1992). Thus these data are intended to represent ALCOVE at peak performance. ALCOVE's fit to the empirical data is highly significant ($R^2 = 0.2119$, $F(1, 74) = 19.89, p = 0.000028$), but much weaker than that of the algebraic complexity (again, $R^2 = 0.5037$, $F(1, 74) = 75.11, p < 0.000001$). Nonetheless, ALCOVE alone explains less than one-quarter of the variance in the human data, while the algebraic theory explained more than one-half.

As mentioned, ALCOVE performance is orthogonal to parity, so one might wonder how much of its poor performance is due to performing equally on Up and Down cases. However, even when ALCOVE is given parity "for free", i.e. parity is entered into the linear regression along with ALCOVE performance, the bilinear ALCOVE-plus-parity model still explains less of the variance ($R^2 = 0.4674$) than does algebraic complexity alone ($R^2 = 0.5037$), though of course the margin is diminished. Hence no matter which way the analysis run, it seems that the learnability of these concepts is better accounted for by their algebraic complexity than by ALCOVE's performance.

6.2. Typicality effects

As discussed above, the algebraic theory extends simply to model the notion of typicality, namely by variations in the degree to which various objects satisfy the regularities in a given power series, captured in the number τ (Eq. (39)). This section investigates the fit of τ to data on individual objects embedded within categories, specifically performance on the individual objects in the (Feldman, 2000) dataset.

[7] This stopping criterion was selected to show ALCOVE in the best light, by creating the greatest possible discrepancies between the various cases, and thus the most information about performance. At earlier points in learning, the various cases have not yet grown very distinct; at later points, they tend to all have converged to the same asymptotic value. Alternatively, as a more faithful model of the true learning circumstances under which the data were collected, ALCOVE's performance can also be evaluated after a period of time that is proportional to the actual learning time available to subjects in the original experiment. This method yields slightly a slightly worse fit to the data, $R^2 = 0.1943$, so the method more favorable to ALCOVE was adopted.

As discussed above, this large dataset includes measurements of human performance for each of 76 logically distinct concept types. All together, these 76 concepts comprise 1072 distinct objects. (As before, we are considering these objects as points in abstract 3- or 4-dimensional Boolean space in canonical coordinates, ignoring the mapping from abstract points to observable features, which was randomly permuted; see Feldman, 2000.) The dataset includes data for each of these 1072 individual objects (never presented in the original paper) with each of the 1072 means drawn from about numerous trials and numerous subjects (the exact numbers depend on condition; see Feldman, 2000). A classical result in the category literature is that classification performance improves with more typical members of a category (Rips, Smith, & Shoben, 1973; Rosch, Simpson, & Miller, 1976), suggesting that proportion correct on a learning task is a reasonable proxy for subjective typicality. Here we consider how well these numbers are predicted by the algebraic quantity τ, which measures how well each object fits into its parent concept's power series.

As discussed above, algebraic typicality τ can be measured with respect to two complementary categories, the one to which the test belonged (i.e. the positive examples on positive trials, the negative examples on negative trials), or the one to which it did not belong (i.e. the positive examples on negative trials, the negative examples on positive trials). These will be referred to as τ_{ispi} and τ_{contra} (because they measure the given object's relationship to the "same side" of the classification boundary and the "other side" of it, respectively). To estimate typicality we use a weighted sum $\hat{\tau}$ of these two τ's,

$$\hat{\tau} = a\tau_{\text{ispi}} + b\tau_{\text{contra}}, \tag{46}$$

with the weights a and b estimated from the data. In the plots below, for clarity of presentation $\hat{\tau}$ has been normalized to the unit interval; this does not affect the statistical analyses in any way.

Algebraic typicality $\hat{\tau}$ and ALCOVE's performance on each object were entered into linear regressions, each separately and then jointly, on human performance with the 1072 datapoints. In the separate regressions, each measure significantly predicted human performance, with $\hat{\tau}$ accounting for about 12% of the variance ($R^2 = 0.1221, F(1,1070) = 148.75$, $p < 0.000001$) and ALCOVE about 7% ($R^2 = 0.0677, F(1,1070) < 0.000001$). (Note that the proportion of variance accounted for by both theories is much lower than in previous analyses because there is much more variance in the 1072-point dataset than in the 76-point dataset; the latter consists of means of the former.) When the two measures are entered into a simultaneous linear regression, the contribution of $\hat{\tau}$ is about three times greater than ALCOVE as measured either by standardized regression coefficient ($\hat{\tau}$: 0.2956; ALCOVE: 0.0979) or by t ($\hat{\tau}$: 8.654; ALCOVE: 2.8659). However, note that even though this analysis discounts any partial correlation between $\hat{\tau}$ and ALCOVE, ALCOVE's own separate contribution is still significant ($p = 0.0042$), though that of $\hat{\tau}$ is more so ($p < 0.000001$). Hence, in summary, algebraic typicality predicts human performance much better than does ALCOVE, but ALCOVE still makes a significant independent contribution. Fig. 12 gives summary plots of human performance vs. each theoretical measure.

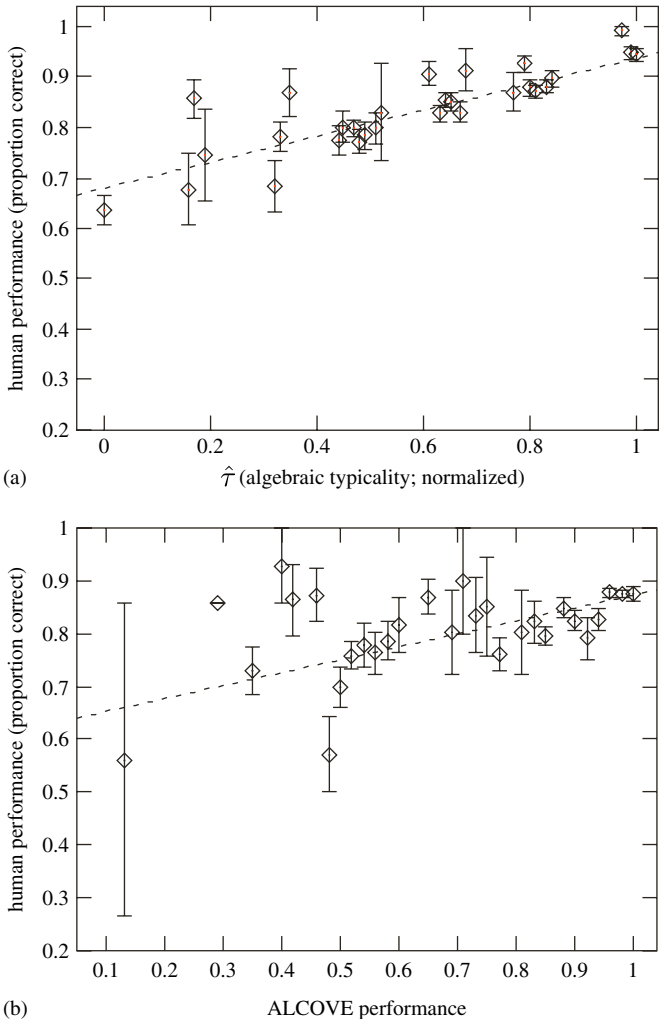


Fig. 12. Proportion correct for the 1072 individual objects (members of the 76 concepts from Figs. 8 and 10; data from Feldman, 2000), plotted against (a) predicted algebraic typicality ($\hat{\tau}$, normalized to the unit interval) and (b) performance by ALCOVE (Kruschke, 1992). In each of these plots the data have been collapsed into 29 distinct means with standard error bars for ease of plotting; however, the statistical analyses reported in the text reflect the 1072 raw numbers. (There are actually 29 distinct levels of $\hat{\tau}$ in the raw dataset, so the ALCOVE scores were rounded into 29 bins to enable a fair visual comparison. Again note though that this binning is reflected *only* in these plots, and not in the analyses themselves.)

The fact that algebraic typicality and ALCOVE both contribute significantly to predicting subjects' performance suggests some kind of hybrid model in which both algebraic decomposition and exemplar storage play roles (see discussion of hybrid models). Along these lines, the joint bilinear prediction of algebraic complexity τ and

ALCOVE explains about 17% of the variance ($R^2 = 0.1733, F(2, 1069) = 112.02, p < 0.000001$) more than either ALCOVE or $\hat{\tau}$ by themselves.

6.3. Summary of the empirical data

In summary, preliminary empirical support for the role of the various algebraic measures in human learning is encouraging.

(i) Algebraic complexity λ^T explains more than twice as much of the variance in mean conceptual difficulty in the large Feldman (2000) dataset as does ALCOVE.
(ii) Algebraic typicality τ explains almost twice as much of the variance in performance on individual objects in the large Feldman (2000) dataset as does ALCOVE.

From these facts, it is reasonable to conclude that human learners' representation of discrete-featured patterns is strongly influenced by the degree of inherent regular structure they exhibit, as captured by the concept algebra or something much like it. As discussed, the data supports a role for exemplar storage as well, although it appears from these data to be a quantitatively smaller one. More detailed investigations of the empirical fit of the algebra to human concept learning, especially in comparison to a modern hybrid model such as SUSTAIN, are deferred to future work.

7. The significance of linearity

As discussed above, some of the data support the hypothesis that observers are biased specifically towards *linear* concepts, that is, concepts with spectral power at degree one or less (the same concepts that are subject to especially simple algebraic decomposition, as outlined in the section above on "an algebra of simple concepts"). Why? Can this preference be justified in any way? A thorough answer to this question might relate the degree K to the objective probability that a concept actually holds in the world, e.g. in a Bayesian framework; see Feldman (2004) for steps in this direction. There are, however, some more immediate arguments.

7.1. Why linear concepts?

First, observe that a reasonable observer considering implicational concepts *must* pick some upper limit of the degree K of concepts to be considered—simply because including the maximum of $K = D - 1$ provides complete expressive power and hence zero inductive power (see discussion of the bias-variance tradeoff above). By contrast, allowing only minimally complex $K = 0$ concepts gives *no* expressive power (or more precisely, allows one to express no *relationships* among variables), which is also undesirable. $K = 0$ is too low and $K = D - 1$ is too high, so the observer must pick something in the middle. Apparently, $K = 1$ is the human choice.[8]

Why $K = 1$? One part of the answer is that linear concepts provide a substantial step up in expressive power from constant ($K = 0$) concepts. Linear concepts, while still very formally simple, are capable of expressing many of the regularities that actually hold in the natural world. This section briefly documents this argument by demonstrating that certain key classes of natural concepts are linear.

7.2. Expressive power of linear concepts

How rich is the set of linear concepts? Can it express rules that actually occur in the real world? This section shows that certain extremely salient and important types of real-world structure fall in the class of linear concepts. This in turn justifies human observers' use of linear structures as their underlying "theory" of the world in Murphy and Medin's sense.

Example 1 (the first "amoeba world") included what was referred to as a *species*. In biology, a species is a set of organisms with a consistent structure, manifesting a large set of properties that are highly correlated with one another. If x is a bird, then x will have wings, feathers, and a beak, etc.—properties that are not logically related to one another, but are contingently related to one another in the extant biological world. In Boolean language, this corresponds to a nexus of pairwise implications[9] of the form $x \to a, x \to b, x \to c$, e.g.

$$\begin{aligned} &\texttt{is\_a\_bird} \to \texttt{has\_wings},\\ &\texttt{is\_a\_bird} \to \texttt{has\_feathers},\\ &\texttt{is\_a\_bird} \to \texttt{has\_beak}. \end{aligned} \tag{47}$$

A set of rules of this form is called a *species*. The "essential" property x itself in this formulation may not be observable, but consistently give rise to observable properties (cf. the "causal features" of Sloman et al., 1998; see also Medin & Ortony, 1989). A theorem follows immediately.

Theorem 2 (*Linearity of species*). *A species is a linear concept.*

(Proof in the appendix.) Hence linear concepts include one of the most basic organizational schemes in the natural world. Needless to say, "species" as defined here need not be restricted to varieties of animals and plants, but also encompasses natural, artifactual, social, and conceptual categories of all kinds.

[8] And the choice underlying Bayes nets, which decompose probability distributions into pairwise dependencies.

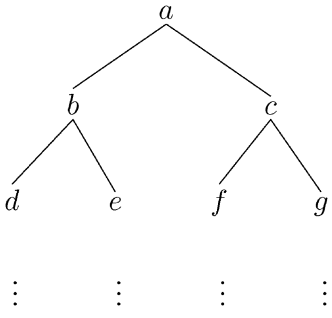
[9] Of course, in a modern view none of these implications would be regarded as logically necessary, though each may be highly probable. Again this means that $x \to a$ may be replaced in practice with $p(a|x) > 1 - \varepsilon$, etc.

If other species do not have any overlapping features, then the implications will also go in the opposite direction. For example, if species x has properties a_x, b_x, (i.e., $x \to a_x, x \to b_x$), and species y has non-overlapping properties a_y, b_y (i.e., $y \to a_y, y \to b_y$), and no other species exists in this world, then a number of other implications immediately follow:

$$\begin{array}{l} a_x \to x \\ b_x \to x \\ a_x \to b_x \\ b_x \to a_x \\ \text{------} \\ a_y \to y \\ b_y \to y \\ a_y \to b_y \\ b_y \to a_y. \end{array} \tag{48}$$

These implications can be grouped into two equivalence classes of implications, the x group and the y group (above and below the dotted line, respectively), within each of which all the properties imply all the others. Obviously, this structure is all linear.

In a more complicated world, not all the features need be unique. In the real world, for example, robins have feathers, but so do sparrows, eagles, and ostriches. These species are all sub-categories of birds, and hence share bird properties, but also have other properties that are unique. Likewise, birds share properties with other classes of animals (they breathe, they eat), but have other properties special to the class. More generally, one imagines an entire cladistic taxonomy of species types, with narrower and narrower subordinate categories towards the bottom, and larger and larger superordinate categories towards the top, e.g.



(49)

A concept of this form (with an arbitrary degree of brachiation at each node and an arbitrary number of levels) is called a *tree of species*. Its structure is potentially far more complicated than that of a single species—indeed this form captures much of the structure of the biological world. Yet it too is linear.

Theorem 3 (*Linearity of trees*). *A tree of species is a linear concept*.

(Proof in the appendix.) Keil (1981) has argued that the tendency to organize categories into trees is a fundamental property of human cognition. Mathematically, because species and trees of species are linear concepts, the tendency to regularize conceptual hierarchies to these simple types can be seen as a special case of the regression to linear concepts.

Again, the argument is not that linear rules can express the entire world; they cannot. Much more complicated structures are possible—scientific theories of how the world works are full of them. As mentioned above, realistically a learner might do well by varying the threshold of degree below which regularities are represented. But these arguments suggest that a truncation specifically at linear concepts retains some particularly important ones, and might thus represent a reasonable default level at which to optimize the tradeoff between inductive leverage and expressivity.

8. Conclusion

In recent years, there has been progress in understanding human concept formation as a "rational" process, that is, one that represents a coherent attempt to estimate some definite property of the environment. This idea has sometimes been realized in terms of Bayesian models (Anderson, 1991; Tenenbaum & Griffiths, 2001), in which the observer seeks to recover the true structure of environmental categories via Bayesian combination of uncertain or stochastic evidence. A common theme in the statistical learning literature is the need to adopt a *model class* as a basis for induction: in effect, a choice of what kinds of "regularities" the observer recognizes. Once a model class is chosen, it becomes possible to estimate the parameters of the model that best explain the data, usually based on a criterion that combines some type of complexity minimization with some measure of fit to the examples. But what is the "model class" of human concept learning? What form does the human learner expect regularities among featural attributes to take? In Murphy and Medin's term, what is the learner's "theory" of the environment?

The argument in this paper is that, for objects described by discrete features, human learners model environmental regularities as simple quasi-causal laws; and form complex representations of the observed pattern via algebraic combinations of these atomic laws. This basic premise leads to a formalism in which concept representations have a certain structural form. The contribution of this paper has been to work out how such algebraic representations could work under certain assumptions about the form of regularities. In the algebraic representation, the observer decomposes the observed pattern of property variation into components of various degrees, from those that are stable and orderly (low degree) to components that are more irregular (high degree), and everything in between. The human bias towards conceptual "simplicity"—a sometimes vague notion—can then be described in mathematical

terms as a bias towards low degree in the concept representation. The empirical evidence suggests that human learning of these kinds of patterns is indeed, impeded in direct correlation with the complexity of the pattern in this representation.

Recently, several theorists have been able to supply a rationale for simplicity principles by showing a connection between complexity minimization and optimization under some inference theory—that is, between simplicity and veridicality (Chater, 1996; Rissanen, 1989). In a sense, the algebraic theory plays a similar role with respect to Boolean minimization: it helps to explain it by showing its connection to inference. Extracting and representing the featural regularities that govern the observations allows the observer to explicitly model what variables are related to what other variables in what ways—a foothold into the causal structure of the environment. Thus the structure of the algebra reflects an unconscious model of what types of rules are liable to operate in the world (cf. Richards' 1988 "principle of natural modes").

A canard among Machine Learning researchers holds that DNF (disjunctions of conjunctions) is the natural representation for *data*, while conjunctive normal form (conjunctions of disjunctions, CNF) is the natural representation for *inferences* about data, e.g. in the mind of an observer. The intuition is that DNF lists all the objects that obey the concept, while the CNF lists all the true statements about those objects. From a psychological perspective this is a bit unsatisfying, because the conjuncts contained in the CNF may vary enormously in subjective salience or compellingness—not all of them are equally likely to play a role in human subjects' generalizations about the data. The algebraic power series improves upon this notion by, in effect, partialing out the conjunctive regularities in the data into components of differing degrees of formal complexity, i.e., factoring the data into the 0th order, 1st order, etc., parts.

It is clear that without further extensions the algebraic approach does not give any account of a number of important aspects of human concept formation, including representation of concepts defined over continuous features, the role of context and background knowledge (Heit, 2001; Pazzani, 1991); and the neural substrate underlying categorization processes (Ashby et al., 1998; Ashby & Ell, 2001; Knowlton & Squire, 1993); and others. Naturally, some understanding of how algebraic ideas might be applied to these questions is desirable, and several ideas have been briefly sketched above. The main aim of this paper, though, is to establish the basic theory of discrete patterns at a "competence" level, and further pursuit of these other issues is deferred to future work. In addition, it should be noted that the theory's ability to fit human data, notwithstanding the encouraging comparison with ALCOVE discussed above, is still largely unknown. Comparisons with more recent accounts such as SUSTAIN are particularly important before any firm conclusions can be drawn.

At best, then, the concept algebra represents only part of the story, even in the limited case of discrete patterns to which it directly applies. Unlike many of the learning models in the literature, it is a pure competence theory—an attempt to characterize at a formal level the abstract function computed by the human learner. As discussed above, there are intriguing connections between the concept algebra and some performance theories. The two types of theories seem to shed light on each other, in that the performance theories articulate how computations proposed in competence theories are actually carried out, while competence theories show why procedures described in performance theories actually "make sense." Thus an important next step for the algebraic approach will be to develop more psychologically realistic models of how algebraic pattern decomposition might actually be carried out in the brain.

Acknowledgments

I am grateful to Cordelia Aitkin, Erica Briscoe, David Fass, Noah Goodman, Fabien Mathy, Whitman Richards, Ed Stabler, and Josh Tenenbaum for many helpful comments, and to Samir Patel for assistance with the ALCOVE analysis. Preparation of this article was supported by NSF SBR-9875175 and NSF SES-0339062.

Appendix A. Proofs of theorems

Proof of Theorem 1 (*Power series representation*). Here we prove Theorem 1 and clarify the sense of "minimal" attached to the power spectrum. Briefly, we impose a (total) ordering on power spectra, and define the power series for $\mathbf{x}$ as any irredundant regularity set equivalent to $\mathbf{x}$ having minimal power spectrum. Thus the power spectrum for $\mathbf{x}$ is unique, and non-unique power series occur when there are multiple power series with identical spectra.

First, we prove the theorem by establishing that there exist irredundant regularity sets equivalent to any object set. Second, we identify an ordering on power spectra and establish that it has a unique minimum.

(i) *Existence*. First, we establish that for any $\mathbf{x}$ there exists at least one set of implication polynomials whose conjunction is equivalent to $\mathbf{x}$. Say there are N objects in $\mathbf{x}$, and thus $2^D - N$ objects not in $\mathbf{x}$. Denote one of these by $y = y_0 y_1 y_2 \cdots y_{D-1}$. The fact that this object is not in $\mathbf{x}$ is equivalent to a rule of degree $D-1$, e.g.

$$y_1 y_2 \cdots y_{D-1} \rightarrow y_0'. \tag{50}$$

$\mathbf{x}$ is equivalent to the conjunction of $2^D - N$ statements of this form—that is, $2^D - N$ rules of maximal degree D. The same is true if each of the D variables is v-valued for some $v > 2$. This establishes that there exists at least one regularity set equivalent to $\mathbf{x}$.

Among such sets, an *irredundant* set is one that has no subsets equivalent to it. (i.e. a regularity set S is called

redundant if there exists $\mathscr{S}' \subsetneq \mathscr{S}$ such that $\wedge\mathscr{S}' = \wedge\mathscr{S}$, and irredundant otherwise.) Loosely speaking, a regularity set is irredundant if one cannot take any of its elements away without breaking its equivalence to **x**. One can create irredundant regularity sets for **x** by beginning with the complete set of regularities obeyed by **x**, and then removing elements one at a time until the equivalence $\wedge\mathscr{S}' = \wedge\mathscr{S}$ no longer holds.

(ii) *Minimality*. We define a total order $\prec$ on power spectra λ (referring to the entire function mapping degree K to power λ^K) by defining $\lambda_1 \prec \lambda_2$ just when λ_1 has less power than λ_2 at some degree K,

$$\lambda_1^K < \lambda_2^K, \tag{51}$$

but agree ($\lambda_1^{K'} = \lambda_2^{K'}$) for all lower degrees $K' < K$. That is, we simply rank power spectra by their magnitude of power, looking first at lower degrees, and breaking ties by looking at higher degrees. Clearly, this is a total order, because regardless of the power at higher degrees, at any one degree K either one spectrum precedes the other or vice versa; and if the two spectra agree at *all* degrees, then they are identical. Hence any two non-identical spectra can be unambiguously ranked, which means that any set of spectra has a unique minimum, including the set corresponding to the set of irredundant power series for a given object set; and this minimal power spectrum is itself unique for a given object set. □

Remark. It is easily seen that the power series that with the minimal power spectrum also usually has *minimal algebraic complexity*, subject to certain constraints on the weight function w_K. Recall that exactly how algebraic complexity is computed depends on the weight w_K assigned to each polynomial of degree K (see Eq. (35)). To see why lesser spectra generally have lower total complexity, observe that any polynomial ϕ^K of degree K can be "replaced" in the power series by n_{K+1} polynomials of degree $K+1$. For example, with Boolean features, the single polynomial a of degree 0 could be replaced by two polynomials of degree 1, namely $b \to a$ and $\neg b \to a$. (Pictorially, the former is a single side of the Boolean square, while each of the latter is one vertex.) However, the latter representation would have greater complexity as long as the weighting function penalizes two degree-1 polynomials more than one degree-0 polynomial. More generally, the weighting function must obey the slope constraint

$$w_K < (n_{K+1})(w_{K+1}), \tag{52}$$

which is obviously satisfied by any increasing function w_K, including the linearly increasing ones considered in this paper. Subject to the constraint in Eq. (52), any lesser spectrum has lower $\sum w_K \lambda^K$ than any greater one, which entails that power series with minimal spectra also have minimal total spectral power. This makes more clear in what sense the algebraic representation constitutes a "minimal" representation of the concept: subject to this very general constraint, the power series is a maximally parsimonious or maximally compressed representation of the concept.

Algorithm. The definition of a minimal power series in terms of spectral precedence suggests an effective and reasonably efficient algorithm for computing the minimal power series. The idea is to add polynomials to the series until the entailed model (that is, the set of objects that obey the power series) matches the object set **x**. Specifically, we consider polynomials in degree order (and in arbitrary order for each fixed level of degree), adding each polynomial to the power series if and only if it (a) is satisfied by **x** and (b) changes the entailed model. When the model matches **x**, we stop, prune any remaining redundant elements, and output the current candidate series.

This algorithm has been implemented in Matlab code publicly available at http://ruccs.rutgers.edu/~jacob/demos/algebra.html. The program (Concept Algebra Toolbox, or CAT) includes a graphical front-end for entering and editing concepts with the mouse. The program plots the power spectrum and computes algebraic complexity for concepts with up to 16 features with arbitrary numbers of values per feature (subject to system-dependent memory and processing constraints).

For noisy data, we can take advantage of the fact that DL($\mathscr{S}$) (from Eq. (44)) can be expected to be a U-shape function of spectral complexity. That is, DL will be high for overly simple power series (which underfit the data) and overly complex ones (which overfit it), and minimal somewhere in between. As the basic algorithm already evaluates candidate power series in spectral order (because it considers regularities one at a time, from lowest degree to highest) we can estimate the power series simply by computing DL for each power series (following Eqs. (40)–(44)), and stopping when the succession of values reaches a local minimum.

Proof of Theorem 2 (*Linearity of species*). Immediate from definitions of *species* and *linear concept*. □

Proof of Theorem 3 (*Linearity of trees*). By construction. Assume

$$x \to a_x, b_x \ldots$$
$$y \to a_y, b_y \ldots$$

are species. Then if

$$x \to y,$$

then x is a subspecies of y, yielding the tree

$$\begin{array}{l} y \\ | \quad \ddots \\ x \end{array}$$

Arbitrarily larger and more complex trees can be constructed in a similar manner. □

Appendix B. The dual algebra

In mathematical logic, the operations $\vee$ and $\wedge$ bear an essentially symmetric relation with each other: for every algebraic statement that holds true, there is a corresponding "inverted" or *dual* statement, in which all the conjunctions and disjunctions have been interchanged, that also holds true (see Davey & Priestley, 1990). For example, the theorem that all formulae have an equivalent DNF (disjunction of conjunctions) has a dual, namely the theorem that all formulae have an equivalent CNF (conjunction of disjunctions).

In the case of the concept algebra, there exists an entire dual algebra whose structure is isomorphic to the one described above, except with conjunction and disjunction swapping roles. Here we briefly describes its properties.

Regularities are defined as before. A *constant* concept ϕ^0 is one congruent to σ, or, in the general n-valued case, congruent to $\sigma(v)$ for some value v. A 1st order concept ϕ^1 is one congruent to

$$\sigma_1 \rightarrow \sigma_2$$

in the Boolean case, or

$$\sigma_1(v_1) \rightarrow \sigma_2(v_2)$$

in the general case (cf. Eqs. (5) and (6)). In general, an implication polynomial of degree K is congruent to

$$\sigma_1 \ldots \sigma_K \rightarrow \sigma_0$$

(Boolean case) or

$$\sigma_1(v_1) \ldots \sigma_K(v_K) \rightarrow \sigma_0(v_0)$$

(general case; cf. Eq. (38)).

The main difference is in how implication polynomials are combined in power series. In the "normal" algebra, each polynomial asserts a rule that all objects in the world obey. In the dual algebra, each polynomial asserts a rule that *some* object in the world obeys. Thus while in the normal algebra the power series $\mathscr{S}(\mathbf{x})$ of an object set $\mathbf{x}$ was defined as a set of polynomials whose *conjunction* is equivalent to $\mathbf{x}$, the dual power series $\mathscr{S}^\delta(\mathbf{x})$ is defined as a set of polynomials whose *disjunction* is equivalent to $\mathbf{x}$:

$$\begin{aligned} \mathscr{W}|_\Sigma = \mathbf{x} &= \Phi^0_{\mathbf{x}} \vee \Phi^1_{\mathbf{x}} \vee \cdots \vee \Phi^{D-1}_{\mathbf{x}} \\ &= \bigvee_{K=0}^{D-1} \Phi^K_{\mathbf{x}} \\ &= \bigvee \mathscr{S}^\delta(\mathbf{x}) \end{aligned}$$

(cf. Eq. (33)).

All the key properties of the original algebra hold in the dual case. Spectra, power, and algebraic complexity are defined exactly as before. Power series are guaranteed to exist and to have a unique minimal power spectrum.

The intuitive motivation for the dual version is different from the normal one, arguably weaker in some senses though perhaps more appropriate in some situations. In any event though the fit of the dual algebra to the human data (e.g. the fit between dual λ^T and the Feldman, 2000 dataset) is poorer than that of the normal algebra. Hence, pending future data, the dual algebra remains a mathematical curiosity, though its potential psychological interest remains to be explored.

References

Ahn, W.-K., Kim, N., Lassaline, M. E., & Dennis, M. J. (2000). Casual status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361–416.

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*, 81–121.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Armstrong, S., Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*, 263–308.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neurophysical theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences*, *5*(5), 204–210.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53.

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, *63*, 516–556.

Barlow, H. (1994). What is the computational goal of the neocortex? In C. Koch, & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 1–22). Cambridge, MA: MIT Press.

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5-4 categories and the category advantage. *Memory and Cognition*, *31*(8), 1293–1301.

Bongard, M. (1970). *Pattern recognition*. New York: Spartan Books.

Bourne, L. E. (1966). *Human conceptual behavior*. Boston: Allyn and Bacon.

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*(6), 546–556.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1960). *A study of thinking*. New York: Wiley.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Sciences*, *7*(1), 19–22.

Davey, B., & Priestley, H. (1990). *Introduction to lattices and order*. Cambridge: Cambridge University Press.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin and Review*, *9*(1), 160–168.

Erné, M. (1993). Distributive laws for concept lattices. *Algebra Universalis*, *30*, 538–580.

Fass, D., & Feldman, J. (2002). Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing*, Vol. 15. Cambridge, MA: MIT Press.

Feldman, J. (1997). Regularity-based perceptual grouping. *Computational Intelligence*, *13*(4), 582–623.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Feldman, J. (2003a). A catalog of Boolean concepts. *Journal of Mathematical Psychology*, *47*(1), 98–112.

Feldman, J. (2003b). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227–232.

Feldman, J. (2004). How surprising is a simple pattern? Quantifying "Eureka!". *Cognition*, *93*, 199–224.

Fodor, J. (1994). Concepts: A potboiler. *Cognition*, *50*, 95–113.

Glymour, C. (2002). *The mind's arrows: Bayes nets and graphical causal models*. Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, *16*, 321–338.

Haygood, R. C. (1963). *Rule and attribute learning as aspects of conceptual behavior*. Unpublished doctoral dissertation, University of Utah.

Haygood, R. C., & Bourne, L. E. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review*, *72*(3), 175–195.

Haygood, R. C., & Devine, J. V. (1967). Effects of composition of the positive category on concept learning. *Journal of Experimental Psychology*, *74*(2), 230–235.

Heit, E. (2001). Background knowledge and models of categorization. In U. Hahn, & M. Ramscar (Eds.), *Similarity and categorization* (pp. 155–178). Oxford: Oxford University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference learning and discovery*. Cambridge, MA: MIT Press.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418–439.

Hovland, C. I. (1952). A "communication analysis" of concept learning. *Psychological Review*, *59*, 461–472.

Hovland, C. I., & Weiss, W. (1953). Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology*, *45*(3), 175–182.

Keil, F. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, *88*(3), 197–227.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 2.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin and Review*, *9*(1), 43–58.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Love, B., Medin, D. L., & Gureckis, T. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, *9*(4), 829–835.

Medin, D. L., Altom, M. W., & Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 333–352.

Medin, D. L., & Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin and Review*, *1*(2), 250–254.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Medin, D. L., & Schaffer, M. M. (1978). Context model of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 355–368.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242–279.

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, *11*, 299–339.

Michotte, A. (1946/1963). *The perception of causality*. New York: Basic Books.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology*, *27*(3), 775–799.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 275–292.

Minda, J. P., Smith, J. D., & Morgan, M. J., Jr. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 659–680.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, *18*(2), 203–226.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Nahinsky, I. D., & Slaymaker, F. L. (1970). Use of negative instances in conjunctive concept identification. *Journal of Experimental Psychology*, *84*(1), 64–68.

Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology*, *64*(6), 640–645.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification recognition and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700–708.

Nosofsky, R. M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory and Cognition*, *19*(2), 131–150.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*(3), 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 416–432.

Pearl, J. (1986). On evidential reasoning in a hierarchy of hypotheses. *Artificial Intelligence*, *28*, 9–15.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Pothos, E. M., & Chater, N. (2001). Categorization by simplicity: A minimum description length approach to unsupervised clustering. In U. Hahn, & M. Ramscar (Eds.), *Similarity and categorization* (pp. 51–72). Oxford: Oxford.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*, 303–343.

Quine, W. (1985). Natural kinds. In H. Kornblith (Ed.), *Naturalizing epistemology*. Cambridge, MA: MIT Press.

Rehder, B. (1999). A causal-model theory of categorization. In M. Hahn, & S. C. Stones (Eds.), *21st annual conference of the cognitive science society* (pp. 595–600). Vancouver.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*, 113–153.

Richards, W. A. (1988). The approach. In W. A. Richards (Ed.), *Natural computation*. Cambridge, MA: MIT Press.

Richards, W. A., & Bobick, A. (1988). Playing twenty questions with nature. In Z. Pylyshyn (Ed.), *Computational processes in human vision: An interdisciplinary perspective* (pp. 3–26). Norwood, NJ: Ablex Publishing Corporation.

Rips, L. J., Smith, E. E., & Shoben, E. J. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*(1), 1–20.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491–502.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Rosch, E. H., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Salmon, W. C. (1998). *Causality and explanation*. New York: Oxford University Press.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1989). Internal representation of universal regularities: A challenge for connectionism. In L. Nadel, L. A. Cooper, P. Culicover, & R. M. Harnish (Eds.), *Neural connections, mental computation* (pp. 104–134). Cambridge, MA: MIT Press.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189–228.

Smith, E. E., & Sloman, S. A. (1994). Similarity- vs. rule-based categorization. *Memory and Cognition*, *22*(4), 377–386.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411–1436.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3–27.

Sober, E. (1975). *Simplicity*. London: Oxford University Press.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Watanabe, S. (1969). *Knowing and guessing: A quantitative study of inference and information*. New York: Wiley.

Watanabe, S. (1985). *Pattern recognition: Human and mechanical*. New York: Wiley.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. I. (1986). Linear separability and concept learning: Context relational properties and concept naturalness. *Cognitive Psychology*, *18*, 158–194.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*, 221–281.