COGNITIVE SCIENCE **19,** 377–405 (1995)

# A Source of Bayesian Priors

DANIEL OSHERSON
*Istituto San Raffaele*
EDWARD E. SMITH
*University of Michigan*
ELDAR SHAFIR
*Princeton University*
ANTOINE GUALTIEROTTI
*University of Lausanne*
KEVIN BIOLSI
*University of Michigan*

Establishing reasonable, prior distributions remains a significant obstacle for the construction of probabilistic expert systems. Human assessment of chance is often relied upon for this purpose, but this has the drawback of being inconsistent with axioms of probability. This article advances a method for extracting a coherent distribution of probability from human judgment. The method is based on a psychological model of probabilistic reasoning, followed by a correction phase using linear programming.

## 1. INTRODUCTION

Human probability judgment has its strengths and weaknesses. Its strength is fecundity, providing reasonable assessments of chance in numerous domains. Its weakness is incoherence, because often, it cannot be represented by numbers in a manner consistent with the probability calculus. The strength has led to the development of expert systems built around a core of human probability assessments (e.g., Andersen, Olesen, Jensen, & Jensen, 1989;

Research support was provided by the Swiss National Science Foundation contract #21-32399.91 to Osherson, by Air Force Office of Scientific Research, Contract No. AFOSR-92-0265 to Smith, and by US Public Health Service Grant #1-R29-MH46885 from NIMH to Shafir. We thank three anonymous reviewers for *Cognitive Science* for constructive and generous criticism of an earlier draft.

Correspondence and requests for reprints should be sent to D. Osherson, D IPSCO, Isituto San Raffaele, Via Olgettina 60, I-20132 Milano, Italy. E-mail: osherson@ratio.hsr.it.

Andreassen, Woldbye, Falck, & Andersen, 1989; Horvitz, Breese, & Henrion, 1988; Long, Naimi, Criscitiello, & Jayes, 1987). The weakness has led to research on the causes and circumstances of incoherent judgment (Shafir, Smith, & Osherson, 1990; Tversky & Kahneman, 1983), and to the elaboration of procedures for inducing better behavior on the part of informants (Kahneman, Slovic, & Tversky, 1982, Part VIII; Winterfeld & Edwards, 1986).

To exploit the strength of human judgment while avoiding its weakness, it would be useful to have a method for minimally revising a person's judgment so as to relieve it of incoherency. Such revision would protect, as much as possible, the insight embodied in the judgment while avoiding conflict with the elementary laws of chance. It is noteworthy that such a method exists for a particular reasoning context, namely, one in which probabilities must be attached to a finite set of statements along with their logical combinations. The method uses linear programming to construct a distribution as close as possible to the numbers that a person proposes for these probabilities. The method works best in the context of a psychological theory that uses a small set of parameters to predict the numbers people choose. Such a theory will be proposed in this article, building on the "Gap Model," developed in Osherson, Smith, Meyers, Shafir, & Stob (1994). The aim of this article is to show how linear programming, in conjunction with the Gap Model, can be exploited to produce coherent probabilities close to human judgment.

The discussion proceeds as follows: Background concepts related to probability and linear programming are reviewed in the next two sections. Section 4 presents the Gap Model, and completes the description of our method for converting raw judgment into probability. Empirical evaluation of the method is contained in Sections 5 through 7. Section 8 is devoted to concluding remarks.

## 2. PROBABILITY IN A FINITE, PROPOSITIONAL ALGEBRA

Our theory bears on probability distributions in finite propositional algebras. This section reviews relevant concepts and definitions, which are illustrated in Appendix I. A more complete treatment can be found in Neapolitan (1990).

Suppose we are given a set $\mathbf{X}$ of $n$ declarative statements $s_1 \ldots s_n$. By a *valuation for* $\mathbf{X}$, it is meant an assignment of truth-value to each of $s_1 \ldots s_n$. A valuation is thus a mapping of $\mathbf{X}$ into $\{true, false\}$, and there are $2^n$ of them. By the *algebra* $\mathcal{A}$ *over* $\mathbf{X}$ is meant the infinite set of propositions that come from combining $s_1 \ldots s_n$ using the logical connectives $\wedge$, $\vee$, $\neg$, and so forth. Under the usual interpretation of connectives, a given valuation imposes a truth-value on each proposition of $\mathcal{A}$.

Now consider a real-valued map **m** defined on the set of valuations for **X** and possessing the following properties:

(1) (a) $\mathbf{m}(v) \in [0, 1]$ for every valuation $v$; and
(b) $\Sigma_v \mathbf{m}(v) = 1$, where $v$ indexes all the valuations for **X**.

We may convert **m** into a probability distribution **P** over $\mathcal{A}$ in the way. For any proposition $\varphi \in \mathcal{A}$, define $\mathbf{P}(\varphi) = \Sigma_w \mathbf{m}(w)$, where $w$ indexes just the valuations that make $\varphi$ true. In this case, **P** is said to be *based on* **m**. Henceforth, by "distribution (over $\mathcal{A}$)" is meant a mapping of $\mathcal{A}$ into reals that is based on some map satisfying (1).

From a given distribution **P**, conditional probabilities are obtained through the familiar equation:

$$(*)\ \mathbf{P}(C \mid B_1 \ldots B_m) = \frac{\mathbf{P}(C \wedge B_1 \wedge \ldots \wedge B_m)}{\mathbf{P}(B_1 \wedge \ldots \wedge B_m)}, \text{ provided that } \mathbf{P}(B_1 \wedge \ldots \wedge B_m) > 0.$$

To be able to move freely between conditional and unconditional probability, we adopt the following terminology: A pair of the form $(C, \{B_1 \ldots B_m\})$, where $C, B_1 \ldots B_m$ are propositions drawn from $\mathcal{A}$, will be called an *argument (of $\mathcal{A}$)*. The statement $C$ is designated the "conclusion" of the argument, whereas the set $\{B_1 \ldots B_m\}$ (which might be empty) is called its "premises." The argument $(C, \emptyset)$ (no premises) is also denoted by $C$. Given distribution **P**, we define $\mathbf{P}(C, \{B_1 \ldots B_m\})$ to be $\mathbf{P}(C)$ if $m = 0$; otherwise, $\mathbf{P}(C, \{B_1 \ldots B_m\})$ is given by the right-hand side of (*).[1]

Suppose that someone assigns probabilities to a subset $S$ of the arguments of $\mathcal{A}$. The assignment can be conceived as a function $H$ that maps $S$ into real numbers. We call $H$ *coherent* just in case it can be extended to a distribution over $\mathcal{A}$. If $H$ is incoherent, we would like to "fix it up" using a method that makes minimal modification to $H$, in some sense. One such method is described in the following Definition and Fact. Suppose that $S$ is a finite subset of the arguments of $\mathcal{A}$, and that $H$ assigns numbers to $S$.

(2) DEFINITION: Let **P** be any distribution over $\mathcal{A}$. **P**'s *error* for $S$ and $H$ is defined as the maximum absolute value of $H(a) - \mathbf{P}(a)$ over all arguments $a \in S$. Any distribution **P** that minimizes the error for $S$ and $H$ is called a *normative envelope* for $S$ and $H$.

(3) FACT: There is at least one normative envelope for $H$ and $S$. Moreover, it can be computed using linear programming.

For an illustration of how to use linear programming for this purpose, see Appendix I and also, Franklin (1980) Section 1.1, Example 6. The use of

[1] For simplicity, we assume in the following discussion that conditional probabilities are well defined, that is, that $\mathbf{P}(B_1 \ldots \wedge B_m) > 0$ for the **P** in question. The assumption is easily lifted in exchange for various provisos in our claims and definitions.

linear programming to find a normative envelope in the sense just described will be called the "*LP* method" for revising a person's probability judgment in view of coherency. The question confronting us is how to make best use of *LP*.[2]

## 3. USING *LP*

We call a set $S$ of arguments over $\mathcal{A}$ *full* (with respect to the algebra $\mathcal{A}$), just in case the following is true: Any coherent mapping of $S$ into the reals can be extended to just one distribution over $\mathcal{A}$. In other words, once numbers are assigned in coherent fashion to a full set of arguments, the arithmetic of probability determines the values of all remaining arguments. The following Definition and Fact describe a full set of arguments that will be important in the sequel. Suppose that algebra $\mathcal{A}$ is generated by statements $\mathbf{X} = \{s_1 \ldots s_n\}$.

(4) DEFINITION: Argument $(C, \{B_1 \ldots B_m\})$ of $\mathcal{A}$ is *elementary* just in case $\{C, B_1 \ldots B_m\} \subseteq \mathbf{X}$. (Thus, an argument is elementary if its premises and conclusion contain no logical connectives, like $\wedge$ and $\neg$.)

(5) FACT: The collection $E$ of elementary arguments is full. Indeed, $E$ is larger than necessary because there are proper subsets of $E$ with only $2^n - 1$ members that are also full. On the other hand, no set of arguments (elementary or not) with fewer than $2^n - 1$ members is full. (The Fact is proved in Appendix II.)

The method *LP* is most plausibly applied to a function $H$ defined on a full set of arguments. For a non-full set, *LP* embodies considerable arbitrariness because linear programming makes arbitrary choices about which normative envelope to impose on $H$, even when the latter is coherent. In view of (5), one way of deploying *LP* may be formulated as follows (where we identify a person with the function $H$ he or she embodies):

(6) SCHEME FOR USING *LP:* Suppose that person $H$ is confronted with a set $\mathbf{X}$ of declarative statements. It is desired to define a (coherent) probability distribution for the algebra generated by $X$ that is close to $H$'s raw intuitions (which may be incoherent). For this purpose, obtain $H$'s probabilities for a full set of elementary arguments of $\mathcal{A}$. Then apply *LP* to compute an associated normative envelope.

---

[2] If normative envelopes are defined in terms of average instead of maximum error, linear programming cannot be applied in a straightforward way. Nonlinear programming (Luenberger, 1984) is a more complicated affair (including problems of local minima), so we have opted for the efficiency, familiarity, and simplicity of the linear programming approach.

Scheme (6) is particularly attractive inasmuch as elementary arguments represent accessible intuitions about chance. This is because only conditioning is in play, rather than logical structure. Consider, for example, the relative ease of judging whether the probability is greater than .5 that human blood contains lithium as an essential component, *assuming that* canine blood does, compared to evaluating the joint probability that *both* types of blood have this property. On the other hand, despite their simplicity, Fact (5) shows that elementary arguments embody sufficient information to determine a probability distribution.

There is only one difficult in applying the scheme. It may be impractical to obtain probabilities from person $H$ for a full set of elementary arguments. For one thing, if the initial statements $\mathbf{X} = \{s_1 \ldots s_n\}$ are numerous, $H$ would be required to make too many judgments. For example, if $n = 10$, then 1023 elementary arguments need evaluation. For another thing, evaluating probabilities might not be as natural for $H$ as other types of judgments. For example, similarity assessments or feature ratings might provide better access to $H$'s knowledge about the domain in question this point is discussed in Szolovits & Pauker (1978).

To remedy these problems and ensure the applicability of *LP*, a psychological theory of elementary arguments is needed. Such a theory would include a relatively small set of parameters which, once set, map the class of elementary arguments into [0, 1]. If the theory is accurate, then for most people there will exist parameter-settings that yield probabilities close to the person's raw intuitions of chance. The probabilities derived from the theory can then be fed to *LP* for rectification. The advantage of applying *LP* to the theory's output rather than directly to judgments of probability, is this. It might be possible to set the parameters of the theory on the basis of smaller, simpler input than a person's judgment about the entire set of elementary arguments. This possibility is illustrated in the experimental work described below. In sum, the existence of a successful theory, **T**, of elementary arguments would allow us to replace Scheme (6) with the following, revised scheme:

(7) REVISED SCHEME FOR USING *LP:* Suppose that person $H$ is confronted with a set **X** of declarative statements. It is desired to define a (coherent) probability distribution for the algebra generated by **X** that is close to $H$'s raw intuitions (which may be incoherent). For this purpose, set the parameters of **T** so that every elementary argument is assigned a probability close to $H$'s judgment. (Use any kind of input from $H$ in order to find the right parameters for **T**.) Then, *LP* to compute an associated normative envelope.

Obviously, Scheme (7) can be implemented only in the context of a theory of elementary arguments. Such a theory is proposed later in this

paper. We then describe experiments designed to determine whether use of Scheme (7) yields a distribution that is reasonably close to a person's original intuitions about probability.

## 4. THE GAP MODEL OF ELEMENTARY ARGUMENTS

We now present a theory of elementary arguments, known as the "Gap Model." Alternative versions of the theory are discussed in Osherson, Shafir et al. (1994) and Smith, Shafir, and Osherson (1993). The present formulation preserves their psychological assumptions but relies on better formulas. Neural net implementation of a related theory is discussed in Sloman (1993). Earlier work on the Gap Model did not include a subsequent correction step to ensure coherence; that is the focus of this investigation.

### 4.1 The Fine Structure of Statements

In the following discussion we limit our attention to statements that have the grammatical form, "subject-predicate," as in:

(8) Bears have three distinct layers of fat tissue surrounding vital organs.

To avoid confusion with "subjects" in the experimental sense, the grammatical subject of a statement will be termed its "object." The object $O$ of (8) is "Bears," the predicate $P$ is "have three distinct layers...," and the entire sentence may be denoted $(O, P)$. Instead of specifying a set of $n \times m$ statements as the basis of our algebra $\mathcal{A}$, it suffices to list $n$ objects and $m$ predicates, with the understanding (as in what follows) that any of the predicates may be applied sensibly to any of the objects; $n \times m$ statements are generated thereby. It is also assumed that the statements are analytically neither true nor false, and exhibit no logical entailments, one to another.

It seems safe to suppose that human judgment of probability is often based on mental representation of the referents of the objects and predicates that compose statements. For example, many people associate large size with both the object and predicate of (8), which renders (8) more probable than the contrasting statement:

(9) Bats have three distinct layers of fat tissue surrounding vital organs.

Such mental representations are no doubt highly structured, perhaps taking the form of "frames" (see Bobrow & Winograd, 1976; Minsky, 1981; Minsky, 1986; Rumelhart & Ortony, 1977; Smith, 1989 for discussion of the psychological reality of frames). However, for simplicity in this study, we assume that mental representations are just nonnegative, real feature-vectors in an appropriate attribute space (see Medin, Altom, Edelson, & Freko, 1982; Osherson, Stern, Wilkie, Stob, & Smith, 1991; Shafir et al., 1990;

Smith, Osherson, Rips, & Keane, 1988; Tversky, 1977; Tversky & Gati, 1982 for similar assumptions in other contexts).[3]

The numbers reflect the perceived degree to which objects possess the corresponding attribute, or in the case of predicates, the perceived degree to which objects satisfying the predicate, typically possess the attribute.[4] No assumption of independence (conceptual or stochastic) is made about attributes. They well may interact (see Malt & Smith, 1984; Medin et al., 1982; Medin & Shoben, 1988) and their values may depend on the totality of objects and predicates in play as discussed in Heit and Rubinstein (in press). More discussion of featural representations of both objects and predicates is available in Osherson, Shafir, and Smith (1994) Sec. 3. Ground-breaking studies include Katz (1972) and Lakoff (1970) and Quillian (1968).

In the presence of feature vectors for each object and predicate, the Gap Model assigns probabilities to every elementary argument. This is achieved through three principles, namely:

(a) a principle that determines the probability of any statement $(O, P)$ considered in isolation
(b) a principle that modifies the features associated with a conclusion in light of the features associated with a single premise
(c) a principle like (b) for multiple premises

We consider these items, in turn. Suppose that human agent $H$ associates feature vectors of length $l$ to every object and predicate. The (nonnegative) value of the $i$th feature associated with object $O$ is denoted by $O(i)$, and similarly, for $P(i)$. Hypothetical attributes and feature vectors are presented in Table 1.

### 4.2 Probabilities of Individual Statements

Let statement $(O, P)$ be given. It is assumed that $H$'s probability for $(O, P)$ varies directly with the overlap in feature content between $O$ and $P$, and indirectly with the feature content present in $P$, but missing in $O$. Overlap is measured using the *minimum* operator, whereas missing material is measured using $\dot{-}$, the "cut-off" operator.[5] For example, according to Table 1, the overlap for ferocity between *horses* and *rage* is $minimum\{horses(3), rage(3)\} = minimum\{3, 11\} = 3$ and the material missing from *horses* is

[3] We exclude negative features in order to simplify our similarity analysis; see formula (13) later.

[4] Thus, we use the term "attribute" for a dimension along which objects and predicates are to be compared, for example, *ferocity* or *size*. We use "feature" for the actual number associated with a given object or predicate on an attribute, for example, 100 for lions on the *ferocity* attribute.

[5] Recall that for all numbers $x$, $y$, $x \dot{-} y = maximum\{0, x-y\}$; hence, $x \dot{-} y$ is subtraction bounded below by 0.

TABLE 1
Hypothetical Feature Vectors
Associated With Three Objects and Two Predicates

| | | objects | | | predicates | |
|---|---|---|---|---|---|---|
| | | *wolves* | *cows* | *horses* | *fight* | *rage* |
| | *attributes* | | | | | |
| 1) | size | 4 | 9 | 8 | 9 | 12 |
| 2) | irritability | 8 | 3 | 6 | 7 | 5 |
| 3) | ferocity | 9 | 2 | 3 | 13 | 11 |

*Note.* *fight*="are more likely to exhibit 'fight' than 'flight' when startled." *rage*="have a brain center for an inborn rage reaction."

*rage*(3) : *horses*(3) = 11 : 3 = 8. The use of cut-off is motivated by the asymmetrical roles of objects and predicates in determining the truth of statements. In particular, the truth of $(O, P)$ seems to depend more on the degree to which $O$ possesses the characteristics demanded by $P$ than vice-versa.[6]

To arrive at probability from the foregoing measures of overlap and disparity, the Gap Model relies on the following rule:

(10) The probability assigned to $(O, P)$ =

$$\frac{\Sigma_{i \leq l}\ minimum\{P(i), O(i)\}}{\Sigma_{i \leq l}[(P(i) : O(i)) + minimum\{P(i), O(i)\}]} = \frac{\Sigma_{i \leq l}\ minimum\{P(i), O(i)\}}{\Sigma_{i \leq l}\ P(i)}$$

The numbers so assigned fall in the interval [0, 1]. They are scale-invariant in the following sense: multiplying all the features by an arbitrary, positive scalar has no effect on the probability attributed to $(O, P)$. Table 1 and formula (10) yield $\frac{8+5+3}{(8+5+3)+(4+0+8)} = .57$ as the probability of (*horses,rage*).

### 4.3 Single Premise Arguments

We now consider the probability assigned to an argument with conclusion $(O, P)$ and sole premise $(O', P')$, as in:

(11) Cows are more likely to exhibit 'fight' than 'flight' when startled. $[(O', P')]$

---

Horses have a brain center for an inborn rage reaction. $[(O, P)]$

In this case, (*horses,rage*) is modified under the impact of (*cows,fight*) prior to the application of rule (10). The impact of (*cows,fight*) may be conceived in the following terms:

[6] Thus, "Corrupt politicans are crooks" commands more assent than "Crooks are corrupt politicians."

Because (*cows,fight*) is a premise, $H$ is requested to assume its truth, which conflicts with the fact that *cows* do not meet $H$'s standards for *fight* with respect to some of the attributes. For example, the shortfall, or "gap," with respect to the irritability attribute is $fight(2) : cows(2) = 7 : 3 = 4$. To resolve the conflict, we assume that $H$ is prepared to increase the value of irritability in objects like *cows*.[7] The relevance of this increase to (*horses,rage*) depends on the similarity of *cows* to *horses* and on the similarity of *fight* to *rage*. High similarity warrants a corresponding increase in *horses*(2), whereas low similarity, with respect to either object or predicate, renders the premise irrelevant to *horses*. We are thus led to the following measure of the impact of an argument's premise upon its conclusion:

(12) Let $(O', P')$ be a premise in an argument with conclusion $(O, P)$. The *impact* of $(O',P')$ on $(O, P)$ with respect to attribute $i$ is defined as: $[P'(i) : O'(i)] \times similarity(O', O) \times similarity(P', P)$.

For the similarity function needed in (12), we distinguish two cases. In the presence of numerous dimensions (e.g., 30 or more), we use the Pearson correlation between the feature vectors associated with each object or predicate; negative correlations are set to zero. Otherwise, we are forced to rely on a cruder measure of the covariation of features, often employed in psychometrics and in other studies of probability judgment (see Gregson, 1975, Section 2.5; Osherson et al., 1991; Stern, 1991). It is defined as the ratio of common to common-plus-distinctive features, which amounts to the following formula when applied to nonnegative vectors $f$, $g$ of length $l$.

$$(13)\quad similarity(f,g) = \frac{\sum_{j \leq l} minimum\{f(j),g(j)\}}{\sum_{j \leq l} maximum\{f(j),g(j)\}}$$

Similarity defined either way is scale-invariant and returns values in [0, 1].[8] To illustrate the use of formula (13) using Table 1:

$$(14)\quad (a)\ similarity(cows,horses) = \frac{8+3+2}{9+6+3} = .72$$

$$(b)\ similarity(fight,rage) = \frac{9+5+11}{12+7+13} = .78$$

[7] The conflict could also be resolved by lowering the predicate values, as implemented in Osherson et al., 1994; Smith et al., 1993). However, (10) must then be reformulated to avoid undesired consequences. The overall model is simpler, as stated here.

[8] It is also symmetric, whereas human similarity judgment is known to violate symmetry in certain circumstances (Tversky, 1977). We have explored "contrast" versions of similarity measures (as in Osherson, 1987; Tversky, 1977) without improvement in the empirical results to be reported below.

(c) impact of (*cows,fight*) on (*horses,rage*) with respect to size = $(9 : 9) \times .72 \times .78 = 0$

(d) impact of (*cows,fight*) on (*horses,rage*) with respect to irritability = $(7 : 3) \times .72 \times .78 = 2.25$.

(e) impact of (*cows,fight*) on (*horses,rage*) with respect to ferocity = $(13 : 2) \times .72 \times .78 = 6.18$

Definition (12) is used in one-premise arguments like (11) as follows: For each feature $i$, $O(i)$ is increased by the impact of $(O', P')$ with respect to $i$. The probability of the argument is then calculated using (10), with the modified $O$ in place of the original. As a result of applying the computations in (14), the feature vector for *horses* in (11) becomes:

$$\begin{bmatrix} 8 \\ 6 \\ 3 \end{bmatrix} + \begin{bmatrix} 0 \\ 2.25 \\ 6.18 \end{bmatrix} = \begin{bmatrix} 8 \\ 8.25 \\ 9.18 \end{bmatrix}$$

The probability given to (*horses,rage*) under the premise (*cows,fight*) is:

$$\frac{8+5+9.18}{(4+0+1.82)+(8+5+9.18)} = .79.$$

The latter probability is higher than computed in Section 4.2 for (*horses,rage*) under no premises. The difference reflects the information carried in (*cows, fight*). More generally, adding premises to elementary arguments only increases the probability of their conclusions, according to the Gap Model. This kind of "monotonicity" is not a general feature of reasoning, as revealed by examples discussed years ago in the context of the "total evidence" requirement see Hempel (1960) Section 2 and references cited there). Even in the limited domain of mammals, monotonicity is sometimes violated in human judgment (see Osherson, Smith, & Wilkie, Lopez, & Shafir, 1990; Smith et al., 1993; Sloman, in press). Such cases are too rare, however, to justify complicating the model here. Nonmonotonicity can be predicted by introducing new assumptions about feature matching as in Sloman (1993), or else by introducing "coverage" variables, as discussed in Osherson et al. (1990).

### 4.4 Multiple Premise Arguments

Now consider multiple premise arguments, such as this one:

(15) Wolves have a brain center for an inborn rage reaction. [$(O'', P)$]
Cows are more likely to exhibit 'fight' than 'flight' when startled. [$(O', P')$]

---

Horses have a brain center for an inborn rage reaction. [$(O, P)$]

The Gap Model assigns probabilities to such arguments through the following maximum concept (which relies on Definition [12], above):

(16) Let $(s, \{s_1 \ldots s_m\})$ be a multipremise argument with conclusion $s$. The *maximum impact* of the premises $s_1 \ldots s_m$ *on* $s$, with respect to attribute $i$, is the largest of the following numbers:

the impact of $s_1$ on $s$ with respect to $i$
$\vdots$
the impact of $s_m$ on $s$ with respect to $i$

For argument (15) and Table 1, arithmetic shows that the maximum impacts are:

(a) 4.16 for size (provided by *(wolves,rage)*
(b) 2.25 for irritability (provided by *(cows,fight)*
(c) 6.18 for ferocity (provided by *(cows,fight)*

Definition (16) is used in multiple premise arguments as follows: For each feature $i$, $O(i)$ is replaced by $O(i)+I$, where $I$ is the maximum impact of the premises on $(O, P)$ with respect to $i$. The probability of the argument is then calculated using (10), with the modified $O$ in place of the original. For example, the feature vector for *horses* in (15) becomes

$$\begin{bmatrix} 8 \\ 6 \\ 3 \end{bmatrix} + \begin{bmatrix} 4.16 \\ 2.25 \\ 6.18 \end{bmatrix} = \begin{bmatrix} 12.16 \\ 8.25 \\ 9.18 \end{bmatrix}$$

and the probability given to *(horses,rage)* under the premises *(wolves,rage)* and *(cows,fight)* is:

$$\frac{12+5+9.18}{(0+0+1.82)+(12+5+9.18)} = .93.$$

Compared to argument (11), the additional premise *(wolves,rage)* raises the probability attributed to *(horses,rage)*. It is evident that the Gap Model's treatment of single-premise and premise-free arguments is a special case of its treatment of multiple-premise arguments.

The foregoing use of maximum is motivated by the following consideration: Suppose that $O_1$ and $O_2$ are highly similar objects, perhaps differing only in name (like porpoises and dolphins). Then an argument of form (a) below should have probability close or identical to that for (b).

$$\text{(a)}\ \frac{(O_2, P)}{(O, P)} \qquad \text{(b)}\ \frac{(O_1, P)}{(O, P)}$$

Such an outcome is obtained by application of the maximum principle. In contrast, summing the impacts of premises gives the wrong result because argument (a) is then assigned an appreciably higher probability than (b) for many choices of predicate $P$. On the other hand, when $O_1$ and $O_2$ are dissimilar we expect (a) to be stronger than (b). It is easy to see that use of the maximum principle assures this outcome whenever the features for $O_1$ do not systematically dominate those for $O_2$.[9] Hybrid models employing a combination of maximum and sum are also possible (see Osherson et al., 1990, p. 199), but in the interests of simplicity, only maximum appears here.

This completes our description of the Gap Model. The reader may verify that the model is insensitive to premise order, and that arguments in which the conclusion appears as premise are uniformly assigned probability 1. These properties of the model have a descriptive appeal, as well as being normatively correct. There is, nonetheles, no guarantee that the "probabilities" generated by the model are coherent; some choices of feature vectors lead to coherence, others to incoherence. It falls upon $LP$ to correct the latter state of affairs, as summarized in Scheme (7) above. The use of the Gap model, followed by application of $LP$ to its outputs, will be denoted "Gap + $LP$" in the remainder of the discussion.

## 5. FIRST EXPERIMENTAL TEST: INPUT FEATURES

Instead of asking a person $H$ to evaluate elementary arguments, we may ask $H$, instead, for the feature-vectors underlying a set of objects and predicates. The Gap Model converts the vectors into probabilities for elementary arguments, which may then be fed to $LP$ for rectification. The resulting distributionn is guaranteed to be coherent. However, it is not guaranteed to approximate $H$'s judgment of probability. Our first experiment was designed to assess the quality of this approximation.

### 5.1 Method

Twenty undergraduates from the University of Michigan participated, recruited by advertisement and paid for their time. There were four parts to the experimental protocol. The students were first presented with a set of objects and predicates. Next, they assigned probabilities to the elementary arguments thus engendered. The same judgments were then made a second time as a reliability check. Finally, participants rated each object and predicate along thirty dimensions. We consider these parts, in turn.

---

[9] For illustration of these points, see Osherson et al., 1990; Smith et al., 1993.

TABLE 2
Sets of Objects and Predicates Available as Options in Experiment 1

| | *Set 1* |
|---|---|
| *Objects:* | Bears, Beavers, Squirrels, Monkeys, Gorillas |
| *Predicate 1:* | have 3 distinct layers of fat tissue surrounding vital organs |
| *Predicate 2:* | have over 80% of their brain surface devoted to neocortex |

| | *Set 2* |
|---|---|
| *Objects:* | Lions, Housecats, Camels, Elephants, Hippos |
| *Predicate 1:* | have a visual system that fully adapts to darkness in less than 5 minutes |
| *Predicate 2:* | have skins more resistant to penetration than most synthetic fibers |

**Presentation of Objects and Predicates**

Subjects were randomly assigned one set of stimuli from the two options shown in Table 2. Each set consisted of five objects and two predicates. Five objects and just one of the predicates yields 5 statements and 80 nontrivial, elementary arguments (i.e., 80 elementary arguments whose conclusion does not figure among the premises). Each of these arguments involves the same predicate in premises (if any) and conclusion. Relying on the same five objects and the second predicate yields another set of 80 elementary arguments of similar character. These two sets of 80 arguments constitute the stimuli delivered to a given participant for evaluation.

**Assignment of Probabilities**

Each person assigned probabilities to his or her 160 arguments, delivered in individualized random order by means of computer. Order of premises was determined randomly for multi-premise arguments. To illustrate, a typical 2-premise argument was presented in the following form:

What is the probability that
    Bears have over 80% of brain surface devoted to neocortex,
given that this property also applies to the following:
    squirrels
    beavers
Probability: _____

The "given that" clause did not appear for 0-premise arguments. Prior instructions emphasized that probabilities were to be assigned while assuming the truth of given premises (if any). Each question was to be treated separately, with no assumptions carried forward. The first two parts of the procedure were performed in immediate succession, and required roughly one hour to complete.

**Reliability Check**

Several days later, participants returned to evaluate all their arguments a second time under a new random order (premise order also was freshly randomized). Previous responses were not made available.

**Feature Ratings**

Participants were asked to rate their five objects and two predicates along thirty dimensions on a scale of 0 to 10. The dimensions used were the same for all participants, and they are shown in Table 3. They were rated in the order listed, first for the two predicates, then for the five objects. These particular dimensions were selected from a set of 80 that figured in earlier experiments (e.g., Osherson et al. 1991). The thirty dimensions in Table 3 received the highest, average rating of relevancy to the objects and predicates appearing in Table 2; the rating was carried out by a separate group of 10 individuals.

### 5.2 Preliminary Analyses

Three participants were dropped from further analyses on the basis of anomalous responses to the probability procedure (e.g., assigning .5 to all arguments, or assigning probability 1.0 to all 0-premise arguments). As a measure of reliability for each of the remaining 17 individuals, a Pearson correlation was calculated between his or her responses in parts 2 and 3 of the procedure. Reliability ranged from .38 to .90, with a median of .68.[10] In all subsequent analyses, we use the average of a participant's two responses to the same argument as his or her "official" judgment about that argument. The coherency of the participants' responses is reported in Section 6.

As explained earlier, each person evaluated two sets of 80 arguments, each set homogeneous in predicate. In subsequent analyses, we keep these sets segregated and thus refer to "half-subjects." Each half-subject evaluated all 80 elementary arguments that arose from the underlying set of five objects and one predicate. The 17 subjects thus represent 34 half-subjects, each analyzed on a within-subject basis.

### 5.3 Performance of the Gap Model With and Without *LP*

Applied to the feature-values associated with a given half-subject, the Gap Model produces probabilities for all 80 of his or her arguments. For each half-subject, we thus compared the mean absolute deviation between the Gap Model's predictions and the value actually assigned, along with the correlation between the latter two numbers. Since each object and predicate was coded along 30 dimensions, the Pearson coefficient was used as a

[10] None of the 17 subjects was dropped on grounds of insufficient reliability. In the next experiment (discussed later), we switched policy and only used data obtained from subjects who passed a reliability threshold.

TABLE 3
Attributes Used in Experiment 1

| | | | |
|---|---|---|---|
| *1* | the color black in visual appearance | *2* | the color gray in visual appearance |
| *3* | being hairy or furry | *4* | having very little noticeable hair or fur |
| *5* | having tough skin | *6* | being big in size and/or mass |
| *7* | being small in size and/or mass | *8* | having a roundish or bulky body shape |
| *9* | having hands | *10* | having long limbs |
| *11* | having molars that are good for chewing | *12* | swimming as a means of transportation |
| *13* | walking as a means of transportation | *14* | being strong |
| *15* | walking and/or standing on hind legs | *16* | walking and/or standing on all fours |
| *17* | being most active at night | *18* | sleeping for extended periods during winter |
| *19* | having high degree of physical coordination | *20* | foraging for food |
| *21* | killing and eating animals | *22* | living in an arctic environment |
| *23* | living in a bush or savannah environment | *24* | living in a plains environment |
| *25* | living in a forest environment | *26* | spending a lot of time on the ground |
| *27* | spending a lot of time in the water | *28* | spending a lot of time in caves |
| *29* | being intelligent | *30* | existing as part of a group |

measure of similarity, with negative correlations set to zero.[11] The median (over 34 half-subjects) absolute deviation is .21. The median correlation is .53 (with 28 positive correlations significant at the .01 level). For each half-subject we next computed normative envelopes for the probabilities offered by the Gap Model. The incoherency of the Gap Model's predictions turned out to be minimal (average error of .007 in the sense of Definition [2]). As a consequence (see Fact [5]), Gap + *LP* makes the same predictions as the Gap Model, alone.

For comparison, we wished to determine the predictive value of a superficial aspect of arguments, and chose the number of premises in an argument for this purpose. Hence, for each half-subject, we computed the Pearson correlation between the number of premises in an argument (ranging from 0–4 over the 80 arguments) and the probability assigned to it. The median coefficient over all 34 half-subjects was .35, hence inferior to use of the Gap Model with features. On the other hand, we also computed for each half-subject the mean absolute deviation between the probabilities assigned to the 80 arguments, and the average of those same probabilities. The median deviation was .17, hence superior to the Gap Model (which does not, of course, peek at the empirically obtained probabilities to make its predictions).

A more sensitive test with input features would tailor the choice of dimensions to the particular subject whose probability judgment is in question. However, even on the basis of preimposed dimensions the Gap Model (with or without *LP*) makes appreciable sense of attributed probabilities, yielding statistically significant correlations in a large majority of cases.

## 6. SECOND EXPERIMENT TEST: INPUT PROBABILITIES

As noted in Section 2, it is often impractical to request evaluation of a full set of elementary arguments. This is because any full set for an algebra based on $n$ initial statements includes at least $2^n - 1$ arguments. It is thus important to determine whether the parameters of the Gap Model can be set accurately on the basis of just a subset of elementary arguments. The model can then be used to generate probabilities for all elementary arguments, with *LP* applied as before.[12]

[11] See Section 4.3. If formula (13) is used instead, the results are slightly inferior to those reported later. On the other hand, in a fifth step to the procedure subjects rated all pairs of mammals for similarity. Using these numbers in place of (13) slightly improves the results. (Since only one predicate figured in a given stimulus set, only its self-similarity is needed for the Gap Model; this was assumed to be 1.)

[12] Some of the data analyzed in this section appeared in Osherson et al. (1994). They are analyzed here in terms of the revised Gap Model, along with the additional step of rectification through *LP*. The data from Experiments 1 and 3 are reported here for the first time.

### 6.1 Method

Fifty-two undergraduates from the University of Michigan completed the first three parts of the procedure described in Section 5.1, that is, everything but feature rating. In this experiment, we insisted on high reliability between the probability judgments given in parts 2 and 3, as measured by the Pearson correlation between them. For 22 participants, this coefficient fell below .70, so their data were dropped from further analyses. The median reliability for the remaining 30 participants was .80. As before, we used the average of a subject's two responses to the same argument as his or her "official" judgment about that argument. Also as before, arguments were segregated by predicate, yielding two half-subjects. Each half-subject evaluated all 80 elementary arguments that arose from an underlying set of five objects and one predicate. Finally, we added the 34 half-subjects from the first experiment to the present data set (ignoring the feature ratings collected for them). The ensuing analyses were thus carried out on a total of 94 half-subjects.[13]

### 6.2 Preliminary Analyses

For each half-subject, we computed a normative envelope for the 80 arguments and associated probabilities. The error of this distribution (according to Definition [2]) is a measure of the incoherency of judgment; zero error implies coherency. Over all 94 half-subjects, the median error was .068, with a minimum of .005 and a maximum of .268. Thus, participants tended to be incoherent.[14]

### 6.3 Performance of the Gap Model With and Without *LP*

Let us now consider the predictive accuracy of our method when supplied with a small set of elementary arguments plus associated probability judgments. The analysis proceeded as follows for each of the 94 half-subjects (analyzed individually):

*First,* ten arguments were randomly selected from the total of 80, to fix the Gap Model's parameters. Different random selections were made for each half-subject. These ten arguments are called the "input" arguments in the following discussion.

*Second,* an iterative procedure was employed to find six vectors of nonnegative numbers, one vector for each of the five objects and one predicate that underlay the 80 elementary arguments. For the length of these vectors,

[13] If the analyses are limited to either (a) the 30 subjects participating in the second experiment, or (b) the 42 subjects in both experiments with intersession reliability of .7 or better, then the performance of the Gap Model is slightly superior to that reported below.

[14] It is possible that our linear programming method is subject to rounding error and that some of our participants were perfectly coherent. We note in this connection that the analysis relied on the widely used MINOS package (Murtagh & Saunders, 1992) for linear optimization.

TABLE 4
Predictive Accuracy of the Gap Model and Gap+*LP* in Experiment 2

| | Arguments | | Predictive Accuracy of the Gap Model | | Predictive Accuracy of Gap+*LP* | |
|---|---|---|---|---|---|---|
| | *# input* | *# predicted* | *Deviation* | *Correlation* | *Deviation* | *Correlation* |
| *1)* | 10 | 70 | .141 | .51 | .157 | .44 |
| *2)* | 20 | 60 | .114 | .69 | .131 | .63 |
| *3)* | 30 | 50 | .093 | .75 | .113 | .72 |

*Note.* The columns headed *Deviation* give the median, average absolute deviation between predicted and observed values. The columns headed *Correlation* give the median correlation coefficient between these numbers. The medians are computed over 94 half-subjects.

we experimented with values of 2, 3, and 4. The resulting analyses yielded virtually identical results; we show only those for 3 attributes.[15] The iterative procedure sought a set of vectors that minimized the average absolute deviation between (a) the probabilities assigned by the half-subject to the ten input arguments, and (b) the probabilities calculated by the Gap Model for the same arguments on the basis of the chosen vectors. The minimization algorithm employed was based on the "direction set" method described in (Press, Flannery, Teukolsky, & Vetterling, 1992, Chapter 10), with a high penalty given to negative numbers. Ten starting points were tried, chosen uniformly-randomly within the unit interval. The best set of features over all ten runs was retained.

*Third,* once the best set of features was obtained in the preceding step, the Gap Model was applied to all 70 elementary arguments *not* participating in the feature-finding stage. A probability was obtained in this way for each. The accuracy of the Gap Model's predictions was measured by calculating the average absolute deviation between its predictions and the half-subject's response for these remaining 70 arguments. The median value of this statistic over all 94 half-subjects is shown in row 1, column 4 of Table 4. We also calculated the Pearson correlation between the probabilities predicted by the Gap Model and those provided by the half-subject, again with respect to the 70 arguments not involved in feature-finding. The median value of this coefficient over all 94 half-subjects is shown in row 1, column 5 of Table 4. *These numbers measure the accuracy of the Gap Model, with no concern for coherence of the predicted probabilities.*

*Fourth,* we computed a normative envelope for the Gap Model's probabilities over all 80 elementary arguments, in the sense of Definition (2). The resulting distribution assigns coherent probabilities to all 70 arguments not participating in the feature-finding stage. Just as for the non-normalized

[15] Because of the low dimensionality of feature vectors (namely, 3), formula (13) of Section 4.3 was used to compute similarity within the Gap Model.

TABLE 5
Predictive Accuracy in Experiment 2 of the "Direct" Method, and of Use of the Mean of the Input Arguments

| | Arguments | | Predictive Accuracy of direct method | | Predictive Accuracy of input means |
|---|---|---|---|---|---|
| | *# input* | *# predicted* | *Deviation* | *Correlation* | *Deviation* |
| *1)* | 10 | 70 | .278 | .14 | .171 |
| *2)* | 20 | 60 | .245 | .31 | .166 |
| *3)* | 30 | 50 | .177 | .55 | .170 |

*Note.* The columns headed *Deviation* gives the median, average absolute deviation between predicted and observed values. The column headed *Correlation* give the median correlation coefficient between these numbers. The medians are computed over 94 half-subjects.

Gap Model, we determined the average absolute deviation between these values and those assigned by the half-subject in question, and we also computed the Pearson correlation between these two sets of 70 numbers. The median values (over all 94 half-subjects) for these statistics are shown in row 1, columns 6 and 7 of Table 4. *These numbers measure the accuracy of Gap + LP. The probabilities put out by the method form a coherent set.*

*Fifth,* as a comparison to the results in Table 4, we computed a normative envelope directly from the 10 input arguments, without the iterative method of Step 2, and with no role for the Gap Model. Applying *LP* in this way to the non-full set of input arguments is called the "direct" method. The resulting distribution was applied, as before, to the remaining 70 arguments, yielding median values for average absolute deviation and for correlation. They are shown in row 1, columns 4 and 5 of Table 5. *These numbers measure the accuracy of the direct method. Again, the probabilities are coherent.*

*Sixth,* as another comparison, we calculated the mean value $m$ of the 10 input arguments, and used $m$ to predict the probabilities assigned to the remaining 70 arguments. For each half-subject, this yields the average absolute deviation of $m$ from the empirically obtained probabilities of the 70 arguments. Its median value over the 94 participants appears in row 1, column 6 of Table 5.

*Seventh,* we repeated steps 1 through 6 above using 20 and then 30 input arguments, in place of 10. The number of predicated arguments thus decreases from 70 to 60 and 50, respectively. The results appear in rows 2 and 3 of Tables 4 and 5.

### 6.4 Discussion

Tables 4 and 5 suggest the following conclusions:

(a) In conjunction with the iterative, feature-finding method described earlier, the Gap Model enjoys reasonable accuracy in predicting the

probabilities assigned to new arguments, starting from those assigned to a small number of input arguments. Compared to using the mean of the latter to predict the former, the Gap Model is 21%, 46%, and 83% more accurate for sets of input arguments of sizes 10, 20, and 30, respectively (using the ratios of the median error). The Gap Model also provides significant information about the relative magnitudes of the probabilities assigned to new arguments, as shown in the correlation coefficients of Table 4.

(b) The normative envelope provided by *LP* is almost as accurate as the uncorrected Gap Model itself. In exchange for coherence, Gap + *LP*'s predictions are only 11%, 15%, and 22% inferior to those of the Gap Model, for sets of input arguments of sizes 10, 20, and 30 arguments, respectively (again, using ratios of median error). The obtained correlations are also only slightly inferior to those obtained from the Gap Model.[16]

(c) Gap + *LP* is considerably more accurate in its predictions than the direct method, which makes no appeal to the Gap Model. (As noted in Section 3, the direct method opens the door to arbitrariness in the choice of the normative envelope selected by linear programming; so, we did not expect it to produce a descriptively accurate distribution.)

In place of medians over subjects, the same conclusions emerge when we consider the number of half-subjects for which one or another method is more predictive. Consider, for example, the mean absolute deviation associated with Gap + *LP* versus the direct method. With 10, 20, and 30 input arguments, Gap + *LP* is more accurate than the direct method for 84, 84, and 79 of the 94 half-subjects, respectively. On the other hand, Gap + *LP* is *less* accurate than the (uncorrected) Gap Model for 57, 75, and 68 of the half-subjects, respectively.

## 7. EXPERIMENT 3: SECOND TEST WITH INPUT PROBABILITIES

The arguments figuring in Experiments 1 and 2 are homogeneous in predicate; the same predicate figures in all the statements composing a given argument. This restriction is lifted in this experiment, which is otherwise similar to Experiment 2.

---

[16] These results are a consequence of the fact that the probabilities offered by the Gap Model are almost coherent, at least for the feature vectors delivered by the iterative procedure. Over all 94 half-subjects, for 10, 20, and 30 input arguments, the Gap Model's median error (in the sense of [2]) is only .011, .012, and .012, respectively.

TABLE 6
Sets of Objects and Predicates Available as Options in Experiment 3

*Set 1*

| | |
|---|---|
| *Objects:* | Bears, Wolverines, Cows, Pigs |
| *Predicate 1:* | Have a brain center that when stimulated gives rise to an inborn rage reaction. |
| *Predicate 2:* | Are more likely to exhibit a 'fight' rather than 'flight' posture when startled. |

*Set 2*

| | |
|---|---|
| *Objects:* | Chimpanzees, Gorillas, Beavers, Squirrels |
| *Predicate 1:* | Have most of their brain surface devoted to neocortex. |
| *Predicate 2:* | Can learn to navigate a complex maze in a matter of minutes. |

### 7.1 Method

Thirty undergraduates from the University of Michigan participated. They were recruited by advertisement and paid for their time.

First, participants were assigned randomly one set of stimuli from the two options shown in Table 6. Each set consisted of four objects and two predicates. They then assigned probabilities to a subset of elementary arguments, now described. Four objects and two predicates gave rise to eight statements and more than a thousand elementary arguments. From among this set, we chose every one- and two-premise argument that met the following condition: at most one premise could have a different predicate than the conclusion, and in this case, the object of the premise and the conclusion were identical. The condition eliminated the more difficult arguments, those in which both the predicate and object vary between a given premise and the conclusion. There were exactly 80 elementary arguments that meet the stated condition (excluding trivial arguments in which the conclusion figures among the premises).

Participants assigned probabilities to their 80 arguments in individualized, random order, by means of the same procedure used in Experiments 1 and 2.

### 7.2 Performance of the Gap Model With and Without *LP*

To evaluate predictive accuracy, we proceeded as in Experiment 2 for each of the 30 students. In summary:

*First,* ten "input" arguments were randomly selected from the total of 80.

*Second,* an iterative procedure was employed to find six vectors of nonnegative numbers, one vector for each of the four objects and two predicates that underlay the 80 elementary arguments. We used a vector length of 3 (the same as used in the second experiment; formula [13] was, therefore, employed for similarity calculations). The iterative procedure sought a set of vectors that would minimize the average absolute deviation between (a) the probabilities assigned by the subject to the ten arguments, and (b) the probabilities calculated by the Gap Model for the same arguments, on the basis of the chosen vectors.

TABLE 7
Predictive Accuracy of the Gap Model and Gap+*LP* in Experiment 3

| | Arguments | | Predictive Accuracy of the Gap Model | | Predictive Accuracy of Gap+*LP* | |
|---|---|---|---|---|---|---|
| | *# input* | *# predicted* | *Deviation* | *Correlation* | *Deviation* | *Correlation* |
| *1)* | 10 | 70 | .150 | .66 | .170 | .58 |
| *2)* | 20 | 60 | .093 | .82 | .129 | .70 |
| *3)* | 30 | 50 | .076 | .87 | .105 | .78 |

*Note.* The columns headed *Deviation* give the median, average absolute deviation between predicted and observed values. The columns headed *Correlation* give the median correlation coefficient between these numbers. The medians are computed over 39 subjects.

*Third,* once the best set of features was obtained in the preceding step, the accuracy of the Gap Model's predictions was measured by calculating the average absolute deviation between its predictions and the participant's response for the remaining 70 arguments. The median value of this statistic for all 30 individuals is shown in Table 7, along with the relevant Pearson correlation.

*Fourth,* we computed a normative envelope for the Gap Model's probabilities over its elementary arguments, in accordance with Definition (2). However, because it was not computationally feasible to use all possible elementary arguments for this purpose, our normative envelope was based on a random sample of 25% of them (different random samples for each participant). As before, the resulting distribution assigned coherent probabilities to all 70 arguments not participating in the feature-finding stage. We determined the average absolute deviation between these values and those assigned by the subject. We also computed the Pearson correlation between these two sets of 70 numbers. The results are shown in Table 7.

*Fifth,* as a comparison to the results in Table 7, we calculated the mean value $m$ of the 10 input arguments, and used $m$ to predict the probabilities assigned to the remaining 70 arguments. See Table 8.[17]

*Sixth,* we repeated steps 1 through 6 above using 20 and then 30 input arguments, in place of 10. See Tables 7 and 8.

### 7.3 Discussion

The tables suggest that the Gap Model, both with and without rectification by *LP*, enjoys considerable predictive accuracy. This finding is reaffirmed by the number of subjects for which Gap and Gap + *LP* are more accurate than the mean of the input arguments. For 10, 20, and 30 input arguments, Gap + *LP* yielded more accurate predictions for 18, 26, and 27 subjects, respectively (out of 30). On the other hand, Gap + *LP* was consistently less accurate than the nonrectified Gap Model, alone.

[17] Due to the large number of valuations (namely, 256) compared to input arguments, the "direct" approach is not a plausible alternative.

TABLE 8
Predictive Accuracy Using the Mean of the Input Arguments, Experiment 3

| | Arguments | | Predictive Accuracy |
|---|---|---|---|
| | *# input* | *# predicted* | *Deviation* |
| *1)* | 10 | 70 | .212 |
| *2)* | 20 | 60 | .208 |
| *3)* | 30 | 50 | .213 |

*Note.* The column headed *Deviation* gives the median, average absolute deviation between predicted and observed values. The medians are computed over 94 half-subjects.

## 8. CONCLUSIONS

Our method Gap + *LP* is only a preliminary attempt to harness the richness of human probability judgment. Despite its rudimentary character, we believe that its predictive accuracy suggests the feasibility of converting a relatively small set of judgments (about features, similarity, probabilities of simple arguments, and so on) into useful measures of chance. Further progress would require many extensions, including the following:

### Domains

- experiments with a wide variety of reasoning domains, instead of just mammals, and with subjects whose expertise in their chosen domain vary

### Linguistic Representation

- extension to a wider set of predicates, for example, to "unbounded" predicates such as *runs faster than deer,* and to "point" predicates such as *runs exactly as fast as deer* (for which the cut-off formula [10] in Section 4.2 is unlikely to be accurate)
- extension to statements having more complicated linguistic structure than object-predicate form

### Mental Representation

- use of more structured representations of knowledge, compared to the attribute-feature system of the Gap Model
- integration of more sophisticated models of similarity, including models that allow the similarity of objects to depend on the predicates with which they are paired (see Cheng, 1991; Heit & Rubinstein, in press; Osherson, Smith, & Shafir, 1986, Section 2.7; Shafir et al., 1990, p. 237; Stern, 1991)

### More Kinds of Data

- extension to other sources on input, e.g., similarities, judgments of "conditional independence" (Geiger & Heckerman, 1991; van der Gaag, 1991), information about default and exceptional properties
- tests of models using nonelementary arguments

### Other Algorithms

- use of alternative means of computing normative envelopes, for example, quadratic instead of linear programming[18]

As a complement to developing methods like Gap+$LP$, it is essential, also, to analyze the kinds of distributions they offer. In exchange for their human-like character, such methods will be limited in the range of chance situations they can model. Useful application thus requires insight into the class of distributions that can be induced on the basis of specified types of input. For example, we have noted already that the probabilities assigned by the Gap Model to arguments are monotonically related to the inclusiveness of the premise-set; adding a premise to an elementary argument cannot lower the probability assigned to the conclusion. It is clear that many distributions fail to enforce this property of arguments, which is why the probability calculus is sometimes considered sufficient to underwrite nonmonotonic reasoning (see Bacchus, 1990; Neufeld, 1989). For another example, call an argument "strong" if it has appreciably higher probability than does its conclusion taken alone. Suppose that the Gap Model rules arguments (a) and (b), below, strong.

$$\text{(a)}\ \frac{(O_1, P_1)}{(O_1, P_2)} \qquad \text{(b)}\ \frac{(O_1, P_2)}{(O_2, P_2)} \qquad \text{(c)}\ \frac{(O_1, P_1)}{(O_2, P_2)}$$

Then it must rule (c) strong as well. For, according to the Gap Model, the strength of (a) and (b) imply high similarity between the two objects and the two predicates, and the strength of (a) implies a sizable "gap" in the premise of (c). This is enough to raise the probability of the conclusion of (c). However, it is easy to find interpretations of the objects and predicates in which the first two arguments are strong, but the third is not.[19]

---

[18] A few of our analyses were carried out using the quadratic programming method devised by Wolfe (see Franklin, 1980, Ch. II.1) with results comparable to those reported here. However, the computational burden of these analyses prevented us from using them throughout.

[19] For example, consider a harmonious couple consisting of an astronaut $O_1$ and his wife $O_2$. Let $P_1$ be, "is selected for a mission to Mars," and let $P_2$ be, "is happy." Then, (a) is strong because of $O_1$'s ambition, and (b) is strong because $O_2$ tends to be happy when $O_1$ is. But (c) is not strong because of $O_2$'s desire to be with $O_1$.

Finally, it is worth emphasizing that the attempt to transform human judgment into coherent, Bayesian priors does not obviate the need to understand the origins of *in*coherence. Indeed, in any situation that leads human intuition to gross violations of the probability axioms, normalization will yield a distribution disconnected from natural judgment and, hence, with no claim to plausibility. Insightful characterization of such situations is thus a precondition for execution of the program we advocate.

## APPENDIX I. ILLUSTRATION OF PROBABILITY CONCEPTS

The illustrations that follow are designed to aid comprehension of Sections 2 and 3.

### Valuations, Algebras, and Distributions

Suppose that there are just 2 members of the initial set $\mathbf{X} = \{s_1, s_2\}$ of declarative statements. For definiteness, they can be imagined to be:

$s_1$ = "It will snow in Altanta during 1999."
$s_2$ = "It will hail in Dallas during 1999."

Then, there are exactly 4 valuations for **X**, namely:

| | | | |
|---|---|---|---|
| $v_1(s_1) = true$ | $v_1(s_2) = true$ | $v_2(s_1) = false$ | $v_2(s_2) = true$ |
| $v_3(s_1) = true$ | $v_3(s_2) = false$ | $v_4(s_1) = false$ | $v_4(s_2) = false$ |

(Think of each valuation as a "possible state of affairs" regarding **X**). The algebra $\mathcal{A}$ over **X** consists of infinitely many statements such as these:

$s_1$ $\quad$ $s_2$ $\quad$ $\neg s_1$ $\quad$ $\neg s_2$ $\quad$ $s_1 \wedge s_2$ $\quad$ $s_2 \vee s_1$
$s_1 \wedge \neg s_2$ $\quad$ $s_1 \wedge \neg s_1$ $\quad$ $\neg (s_1 \vee s_1)$ $\quad$ $\neg(s_1 \wedge s_1)$ $\quad$ $\neg\neg(s_1 \rightarrow s_2)$ $\quad$ $s_2 \vee (s_1 \vee s_2)$

Each valuation imposes a truth-value on every formula of $\mathcal{A}$. For example:

$v_2(\neg s_1) = true$ $\quad$ $v_2(s_1 \wedge s_2) = false$ $\quad$ $v_2(s_2 \vee s_1) = true$
$v_2(s_1 \wedge \neg s_2) = false$ $\quad$ $v_2(s_2 \wedge \neg s_1) = true$ $\quad$ $v_2(\neg\neg(s_1 \rightarrow s_2)) = true$

Now consider the real-valued map **m** defined on the four valuations as follows:

$$\mathbf{m}(v_1) = .1 \qquad \mathbf{m}(v_2) = .2 \qquad \mathbf{m}(v_3) = .3 \qquad \mathbf{m}(v_4) = .4$$

(Thus, **m** gives the probability of each, possible state of affairs.) Since **m** satisfies the conditions in (1), it may be converted into a distribution **P** over

$\mathcal{A}$, as indicated in Section 2. For example, consider the formula, $s_2 \vee s_1$. It is easy to verify that just the valuations of $v_1$, $v_2$, $v_3$ render $s_2 \vee s_1$ true. Hence, $\mathbf{P}(s_2 \vee s_1) = \mathbf{m}(v_1) + \mathbf{m}(v_2) + \mathbf{m}(v_3) = .6$. As an illustration of conditional probability, we have:

$$\mathbf{P}(\neg s_1 \mid s_2 \vee s_1) = \frac{\mathbf{P}(\neg s_1 \wedge (s_2 \vee s_1))}{\mathbf{P}(s_2 \vee s_1)} = \frac{\mathbf{m}(v_2)}{\mathbf{m}(v_1) + \mathbf{m}(v_2) + \mathbf{m}(v_3)} = \frac{.2}{.6} = \frac{1}{3}.$$

## Coherence Through Linear Programming

Here are some arguments of $\mathcal{A}$:

$$(\neg s_1, \{s_2 \vee s_1\}) \qquad (s_1 \vee \neg s_2, \{\neg s_1, s_2\}) \qquad (s_2 \vee \neg s_1, \emptyset) \qquad s_2 \vee \neg s_1$$

The probability according to $\mathbf{P}$ of $(\neg s_1, \{s_2 \vee s_1\})$ was calculated earlier to be 1/3.

Let $S$ be the subset $\{s_1, s_1 \wedge s_2\}$ of (zero-premise) arguments of $\mathcal{A}$. Suppose that $H$ is defined on $S$ so that $H(s_1) = .4$ and $H(s_1 \wedge s_2) = .1$. Then $H$ is coherent because it can be extended to a distribution over $\mathcal{A}$. For example, $\mathbf{P}$ as defined previously is such a distribution. By contrast, if $H(s_1) = .4$ and $H(s_1 \wedge s_2) = .5$, then $H$ is incoherent because no distribution assigns greater probability to $s_1 \wedge s_2$ than to $s_1$. Because $s_1$ is made true by $\{v_1, v_3\}$, and $s_1 \wedge s_2$ by $\{v_1\}$, a normative envelope for the incoherent version of $H$ is found by minimizing the variable $c$ relative to the following linear constraints see Franklin (1980) for discussion.

$$c + v_1 + v_3 \geq .4 \qquad c + v_1 \geq .5 \qquad -c + v_1 + v_3 \leq .4 \qquad -c + v_1 \leq .5$$
$$v_1 \geq 0.0 \qquad v_2 \geq 0.0 \qquad v_3 \geq 0.0 \qquad v_4 \geq 0.0$$
$$c \geq 0.0 \qquad v_1 + v_2 + v_3 + v_4 = 1.0$$

One normative envelope that emerges from the minimization is:

$$\mathbf{m}(v_1) = .45 \qquad \mathbf{m}(v_2) = .275 \qquad \mathbf{m}(v_3) = 0.00 \qquad \mathbf{m}(v_4) = 0.275$$

Its error with respect to $S$ and $H$ is .05, which is the value assigned to $c$ by the minimization. A slight variant of this procedure computes normative errors for arguments with non-empty sets of premises.

## Elementary Arguments

Here are all the elementary arguments of $\mathcal{A}$, except for the trivial ones in which the premises include the conclusion.

$$s_1 \qquad s_2 \qquad (s_1, \{s_2\}) \qquad (s_2, \{s_1\})$$

It can be shown that the first three arguments constitute a full set (as do the first, second, and fourth). Thus, any coherent assignment of probabilities to the first three arguments can be extended to a distribution over $\mathcal{A}$ in just one way. For example, let $H$ be such that $H(s_1) = .6$, $H(s_2) = .5$, and $H(s_1,$

$\{s_2\}) = .7$. Then $H$ is coherent and can be extended only to the distribution based on:

$$\mathbf{m}(v_1) = .35 \qquad \mathbf{m}(v_2) = .15 \qquad \mathbf{m}(v_3) = .25 \qquad \mathbf{m}(v_4) = 0.25$$

## APPENDIX II: SKETCH OF THE PROOF OF FACT (5)

By a "positive conjunction," we mean any formula of the form $s_1 \wedge \ldots \wedge s_j$, where each $s_i$ is an elementary (unnegated) statement. It is easy to verify that $r$ statements lead to $2^r - 1$ logically distinct positive conjunctions of length 1 to $r$. By reflecting on the Venn diagram associated with a distribution over $r$ events, it is clear that the probability of positive conjunctions determines the probability of every formula (because the labeled regions of a Venn diagram are just the positive conjunctions). Hence, the set of positive conjunctions is full.[20] It is thus sufficient to show that the probabilities of any positive conjunction can be deduced from those of elementary arguments. To see that this is true, consider the positive conjunctions $s_i$ and $s_i \wedge s_j$. Their probabilities follow from those attached to the elementary arguments $s_i$ and $(s_j, \{s_i\})$. For $s_i$, this is trivial. For $s_i \wedge s_j$, notice that $P(s_j, \{s_i\}) \times P(s_i) = [P(s_i \wedge s_j)/P(s_i)] \times P(s_i) = P(s_i \wedge s_j)$.[21] For positive conjunctions with more than two conjuncts, it is clear how to proceed by induction. In fact, only $2^r - 1$ elementary arguments are necessary to deduce all of the positive conjunctions in this manner. Fewer than $2^r - 1$ arguments (even in the presence of the sum-to-one constraint) are insufficient because the dimensionality of the space of valuations is $2^r$.

## REFERENCES

Andersen, S., Olesen, K., Jensen, F. V., & F. Jensen, F. (1989). Hugin—a shell for building Bayesian belief universes for expert systems. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence.*

Andreassen, S., Woldbye, M., Falck, B., & Andersen, S. (1989). Munin—a causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence.*

Bacchus, F. (1990). *Representing and reasoning with probabilistic knowledge.* Cambridge, MA: MIT Press.

[20] A more rigorous proof of this step in the argument considers the zero-one matrix relating positive conjunctions to the valuations satisfying them, along with a row of 1's to represent the empty conjunction (equivalently, the sum-to-one constraint). This matrix is row-equivalent to a triangular matrix with 1's along the diagonal. It thus has nonzero determinant, and so, is invertible. This suffices to show that any coherent set of probabilities for the positive conjunctions determines probabilities for all the valuations.

[21] It may be assumed that $P(s_j, \{s_i\})$ is well defined because the coherent mappings $M$ invoked in the definition of "full" send each argument to a finite real.

Bobrow, D., & Winograd, T. (1976). An overview of KRL, a knowledge representation language. *Cognitive Science, 1*(1), 3–46.

Cheng, Y. (1991). Context-dependent similarity. In P. Bonissone (Ed.), *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence* (pp. 41–47). New York: Elsevier.

Franklin, J. (1980). *Methods of mathematical economics.* New York: Springer-Verlag.

Geiger, D., & Hackerman, D. (1991). Separable and transitive graphoids. In P.B. (Ed.), *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence* (pp. 65–87). New York: Elsevier.

Gregson, R. (1975). *Psychometrics of similarity.* New York: Academic.

Heit, E., & Rubinstein, B. (in press). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Hempel, C. G. (1960). Inductive inconsistencies. *Synthese, 12*(1), 28, 439–469.

Horvitz, E., Breese, J., & Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning, 2,* 247–302.

Kahneman, D., Slovic, P., & Tverskky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Katz, J. (1972). *Semantic theory.* New York: Harper & Row.

Lakoff, G. (1970). Linguistics and natural logic. *Synthese, 22,* 151–271.

Long, W., Naimi, S., Criscitiello, M., & Jayes, R. (1987). The development and use of a causal model for reasoning about heart failure. *IEEE symposium on computer applications in medical care* (pp. 30–36).

Luenberger, D. (1984). *Linear and nonlinear programming.* Reading, MA: Addison-Wesley.

Malt, B., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250–269.

Medin, D., Altom, M., Edelson, S., & Freko, D. (1982). Correlated symptoms and simulated medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 37–50.

Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20,* 158–190.

Minsky, M. (1981). A framework for representing knowledge. In J. Haugeland (Ed.), *Mind design.* Cambridge, MA: MIT Press.

Minsky, M. (1986). *The society of mind.* New York: Simon & Schuster.

Murtagh, B., & Saunders, M. (1992). Minos 5.4. Systems Optimization Laboratory, Stanford University.

Neapolitan, R. (1990). *Probabilistic reasoning in expert systems.* New York: Wiley.

Nuefeld, E. (1989). Defaults and probabilities: extensions and coherence. In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning.*

Osherson, D. (1987). New axioms for the contrast model of similarity. *Journal of Mathematical Psychology, 31*(1), 93–103.

Osherson, D., Shafir, E., & Smith, E. E. (1994). Extracting the coherent core of human probability judgment. *Cognition, 50,* 299–313.

Osherson, D., Smith, E. E., Meyers, T. S., Shafir, E., & Stob, M. (1994). Extrapolating human probability judgment. *Theory and Decision, 36,* 103–129.

Osherson, D., Smith, E. E., & Shafir, E. (1986). Some origins of belief. *Cognition, 24,* 197–224.

Osherson, D., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category based induction. *Psychological Review, 97,* 185–200.

Osherson, D., Stern, J., Wilkie, O., Stob, M., & Smith, E. (1991). Default probability. *Cognitive Science, 15,* 251–270.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. (1992). *Numerical recipes in c,* (2nd ed.). New York: Cambridge University Press.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 216–260). Cambridge, MA: MIT Press.

Rumelhart, D., & Ortony, A. (1977). The representation of knowledge in memory. In R. Anderson, R. Spiro, & Montague, W. (Eds.), *Schooling and the acquisition of knowledge.* Hillsdale, NJ: Erlbaum.

Shafir, E., Smith, E., & Osherson, D. (1990). Typicality and reasoning fallacies. *Memory and Cognition, 18*(3), 229–239.

Sloman, S. A. (1993). Feature based induction. *Cognitive Psychology, 25*(2), 231–280.

Sloman, S. A. (in press). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition.*

Smith, E. E. (1989). Concepts and induction. In M. Posner (Ed.), *Foundations of Cognitive Science* (pp. 502–526). Cambridge, MA: MIT Press.

Smith, E. E., Osherson, D., Rips, L., & Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science, 12,* 485–527.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility, and judgments of probability. *Cognition, 49,* 67–96.

Stern, J. (1991). Default reasoning about probability. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

Szolovits, P., & Pauker, S. (1978). Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence, 11,* 115–144.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–362.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89*(2), 123–154.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315.

van der Gaag, L. (1991). Computing probability intervals under independency constraints. In P. Bonissone (Ed.), *Proceedings of the Sixth Workshop on Uncertainty in Artificial Intelligence* (pp. 457–467). New York: Elsevier.

Winterfeld, D. V., & Edwards, W. (1986). *Decision analysis and behavioral research.* New York: Cambridge University Press.