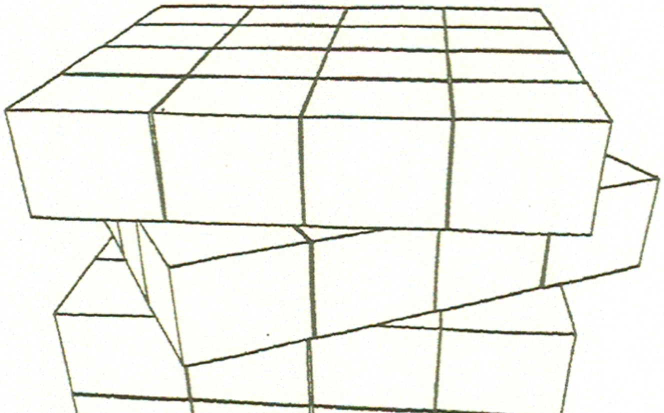
MIT/LCS/TM-232

PLY:
A SYSTEM OF PLAUSIBILITY INFERENCE
WITH A
PROBABILISTIC BASIS



Alexander Yeh
December I982

Abstract:

An overview is given of a system of plausibility inference that will be developed for use in planning. This system, to be called PLY, will be specifically designed to work with propositions of the form "when A is true (occurs), B is likely to be true (to occur)". Previous systems performing similar functions have been designed as aids for such tasks as medical diagnosis (MYCIN and others) and mineral prospecting (PROSPECTOR).

PLY will have a probabilistic basis. Intuitive assumptions to deal with knowledge not explicitly given to the system will be made with the aid of an information-theoretic measure on the amount of information in a probability distribution. Unlike many other systems, PLY will not use these assumptions when the given knowledge indicates they are not tenable. In addition to standard probabilities, PLY will be able to make use of knowledge (information) in the form of correlations and increased/decreased likelihoods, which most people find easier to estimate than probabilities.

PLY's knowledge will be in an organized and structured form, which will help in knowledge acquisition and revision, facilitate system explanations, and lower the storage requirements of the system.

Keywords:

expert system, plausibility inference, probabilistic inference, planning aid

Acknowledgments:

Many thanks are given to Lowell Hawkinson for his guidance and advice on this paper. Thanks are also given to Peter Szolovits for identifying many sources that were useful in preparing this paper. This research was supported by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under contract number N00014-75-C-0661.

This technical memorandum is a proposal for a master's thesis to be supervised by Lowell Hawkinson.

Cover design: Rubik's Revenge by John G. Aspinall.

Table of Contents

1. Introduction

This thesis will introduce a system of plausibility inference called PLY that is to be used as a planning aid. PLY will take in such information as "when statement A is true (event A occurs), statement B is likely to be true", and then use that information, along with information provided by the user as to what is known to be likely to be true, to draw inferences about what other statements are likely to be true.

PLY is a foundation for *expert systems*. That is, PLY may be set up to answer certain questions in an area or field using the knowledge of one or more "experts" in that area, and is then used by others needing specific advice. The types of uses for PLY in planning are similar to the ones for expert systems developed for other areas, such as medicine [Szolovits 78, Kulikowski 79, Shortliffe 79]. A person wanting advice from PLY would go up to it and ask a question. PLY might then elicit additional information (often from the person) before making a reply.

1.1 The Nature of Knowledge in PLY

The relationships that this system will deal with are mostly probabilistic. Furthermore, many of these reflect the opinions of certain people. Given the nature of these relationships, one might wonder, why try to deal with them? Why not just deal with relationships that are deterministic and well-established (such as known cause and effect relationships where the effect will always appear when the cause is present)? Well, the fact is that many important relationships in the world are probabilistic. For example, studies indicate that cigarette-smoking promotes cancer. These studies do not show that every person who ever has smoked a cigarette will get cancer, but that a cigarette smoker has a greater probability of getting cancer than a similar person who does not smoke cigarettes. Although these studies indicate that this relationship between cigarette-smoking and cancer is just probabilistic (as opposed to holding in every case), many people think that this relationship is important enough to put up advertisements telling people not to smoke cigarettes, and to have a warning printed on each pack of cigarettes and each printed cigarette advertisement.

Furthermore, information on the existence and strength of these probabilistic relationships are often not even available as the results of statistical studies (as in the example above), but only as the opinion of one or more 'experts' or people with some familiarity with the subject at hand. In [Shortliffe 75], Shortliffe and Buchanan mentions that this is the case with a good deal of the knowledge used in many areas of medicine. They found that statistical data in the area was very hard to obtain, and when it was obtained, it was often inaccurate. As a result, to get information on the area, they resorted to asking experts in the area for their opinions. Such opinions (in this case, relationships between symptoms, test results, and various ailments), can be: (a) the result of past observations; (b) common 'wisdom' in the field about similar or analogous situations; and/or (c) guesses, hunches, or estimations based on intuition.

1.2 Problems in Past Systems

There have been several problems in past systems dealing with plausibility inference, one of these is that they often need unconditional and conditional probabilities of event occurrences, which people have a hard time estimating. This estimation problem has been encountered in both the PROSPECTOR project [Hart 77], which deals with mineral exploration consulting, and the MYCIN project [Shortliffe 75], which deals with consultations on bacterial infections.

Belief measures which people seem better at handling and estimating than unconditional and conditional probabilities have been proposed. However, the methods of combining evidence used in these measures have no formal basis, making them hard to analyze and hard to alter to take care of cases where they fail (except on an ad hoc basis). Examples of this include the combination methods used in the confirmation measure used in MYCIN [Shortliffe 75], and Dempster's rule of combination (orthogonal sum), 'which is advocated in Shafer's theory of evidence [Shafer 76]. The combination methods used for these measures have been justified on purely intuitive grounds or less.

Another problem is that the assumption of independence and/or conditional independence between event occurrences is often made where in fact it should not be made. Two events are independent if knowledge of one's likelihood does not affect knowledge of the other's likelihood. Two events are conditionally independent on a third event if the former two events are independent when the third event is known to have occurred. These assumptions simplify calculation and estimation problems in systems using a probabilistic basis, and is the only case treated in the evidence combination method of one of the alternative measures (Dempster's orthogonal sum). This problem is widely cited with respect to systems having a probabilistic basis, having been mentioned in, among other papers, [Szolovits 78, Duda 79, Shortliffe 79, Pednault 81].

Often a system in this area is built so that if an assumption that it makes turns out to be a bad one with a few of the events, there will be no clean way (or no way at all) of taking back that assumption for those events. Assumptions that have been made include the independence and conditional independence of event occurrences mentioned above, and also others on the value of certain joint probabilities given the values of their marginals (an example of a joint probability and its marginals is P(A,B), and P(A) and P(B) respectively).

1.3 A Probabilistic Basis

The problems mentioned above can be avoided by adopting a probabilistic basis for a system of plausibility inference. This is because a probabilistic basis can form a formal, consistent framework from which to handle plausibility inferences.

Such a basis can help make explicit the assumptions made concerning the situation at hand, and provide a basis for altering or retracting these assumptions if they turn out to be erroneous. This is important because in just about any system in which one makes a best estimate or approximation, one will have to make assumptions. Also, one will often have to make assumptions in order to make the number of possibilities manageable. However, any assumption can be a bad one to make at certain places, so one should be able to retract such an assumption at those certain places. This entails both being able to easily identify the assumption made that's bad, and then being able to retract or revise it. If one takes the position that no assumptions should be made, then one can run into the problems that will be mentioned later on just using bounds on value ranges and only giving these as final results, and also on the combinatorial explosion of possibilities that one will have to consider.

Another advantage of such a basis is that it can make clear which situation descriptions are possible (the description contains no inconsistencies) and which are not.

One potential problem with using a probabilistic basis is that the conditional and unconditional probabilities that are normally associated with such a basis have the previously mentioned problem of being hard for people to estimate. To get around this problem, one should use alternative measures which have a probabilistic basis, but which people seem to find easier to estimate, when asking people for information. Two such measures are a variant of the correlation coefficient measure used in standard probability and statistics and the confirmation measure used in MYCIN [Shortliffe 75]). Among other things, these two measures have in common the idea of trying to measure the relative "strength" of a relationship between two events, as opposed to the absolute numeric odds involved.

1.4 Applications to Planning

In planning, PLY can be used as a decision-making aid to give advice on such questions as: "Which of my products/services is this potential customer likely to want?", "Which of my goods should I expand my production or inventory of?", "Would this person be suited for the job?", or "What were the likely causes of failure for that plan? (out of a list of possibilities)". Such specific consulting aids can be set up by one or more "experts" in the area(s) in question, and then used by others needing specific advice when the "experts" are gone, busy, or less cost-effective.

Also, PLY can be used as a feedback aid on trying to develop criteria for answering such questions. Showing the experts how PLY reacts to various situations helps them understand implications of the information they gave to it.

In addition, PLY can be used as a teaching aid to help people see more specifically the opinions of experts on the relative importance of various factors in an area in two ways: First, by showing the organization and content of the information that the experts gave to PLY (information which by being forced to be acceptable to PLY is hopefully fairly unambiguous, and which presumably would have been tested on PLY), and second, by showing how PLY would react to typical situations using the information of the experts.

In a fairly heavily automated planning environment, PLY could also be used to make automatic warning or reminder messages when certain conditions arise. For example, if PLY determined that a high possibility exists that the inventory level to be maintained for a certain product should increase soon, it could issue a message to notify the appropriate people.

2. Survey of Existing Approaches and Systems

In this chapter, a survey is given of some existing systems and approaches in this area, and some of their shortcomings and features are indicated.

2.1 Bayesian (Standard Probability)

A standard method of dealing with uncertainties is to use classical or standard probability [Szolovits ??, Szolovits 78, Raiffa 68], assigning to each possible event or statement the probability of it occurring or being true. With probability, the effects of knowing that certain statements are true on the probabilities of the other statements being true are reflected in conditional probabilities. A common use for conditional probability is to give the probability of a hypothesis given some evidence (P(H|E), where H is a hypothesis and E is some evidence).

Standard probability has a nice, formal, mathematical basis. Unfortunately, it also has many problems when directly used in a system of plausibility inference. For one thing, many people have a hard time estimating probabilities. Developers of both the PROSPECTOR [Hart 77] and MYCIN [Shortliffe 75] projects found that people they tried to get probability estimates from could not generally give accurate estimates. Part of this problem is due to the fact that often a person just does not know the chances of a particular event occurring, so that any guess (s)he gives on the probability will be just that, a guess. As will be discussed in a later section, Shafer develops a theory to deal with this in [Shafer 76]. Another problem is the large amount of data that one has to collect for this method. Given k distinct events one will have to find the 2^k probabilities of the possible combinations of those events occurring or not occurring. For any non-trivial k, this will be an enormous number.

To avoid having to ask people for the probability of each combination occurring, designers of existing systems have assumed varying degrees of independence between events and often accompanied this with the use of Bayes' rule. However, a problem with most independence assumptions is that the only assumption that can be made is that of total independence between two groups of events. This is avoided by using a method found in [Duda 79] in the appendix by Kurt Konolige, which assumes as much independence as needed to fill in gaps in the information given. Unfortunately, this method will also try to find the probability of every combination of the events occurring or not occurring, so that for k distinct events, it will try to find all 2^k combination probabilities.

A classical probabilistic formalism also has the deficiency that it does not tell one how to organize the knowledge in a system of plausibility inference so that people checking the system will be able to find out easily what knowledge it may be missing, or what erroneous information it might have.

2.2 Decision Analysis

Decision analysis, as advocated in [Raiffa 68], uses probabilities and the idea of the costs of outcomes for a method of deciding what alternative to take out of a possible set of alternatives. Raiffa gets around the problem of people making bad probability estimations, by showing in [Raiffa 68] that when one has to make a decision and one is trying to minimize the expected subjective costs (utilities), then one should treat a guessed-at subjective probability no differently than a probability based on large amounts of statistics. So, when this situation holds, the inability to make good probability estimations should not be a real liability.

A major problem with this analysis is that often a problem cannot be stated in terms of minimizing expected costs, for costs are often very hard to pin down. This is especially true of irreversible events, events for which no amount of money or resources spent will reverse their effects. For example, what is the cost of a human life, or of being permanently crippled? A related problem is that because of the inability to reverse the effects of any of these events, their associated cost will vary considerably in the minds of different people, with much of that variation probably depending on whether that person is affected by the event in question or not. So, the question arises of who should determine the costs to be used in the calculations. Another problem with costs is that [Kahneman 82] shows that people often have trouble maintaining a consistent cost from one situation to the next, even when the same numeric gain or loss is involved, by showing how the wording of a description of a situation can influence the decision that the people make.

Decision analysis also has the problem of structuring a problem as a sequence of decisions to be made. As mentioned in [Pearl 82], people often do not think of a problem in these terms, but rather in terms of a group of interrelated factors. This can make it hard to set-up a problem for decision analysis.

2.3 PROSPECTOR

PROSPECTOR [Hart 77, Duda 79] is a computer system designed to aid mineral exploration. It takes in user information and then tells what and where minerals are likely to be found. PROSPECTOR combines a probabilistic basis with the idea of an *inference net*, a network to structure and organize the hypotheses and evidence. The rational for this is to combine the advantages of the formal, mathematical basis of the former with the ability given by the latter to more easily acquire and examine knowledge from various experts. Also, the latter helps to limit the possible inferences that the system needs to consider.

To get around the problem of the inability of most people to make accurate probability estimates, PROSPECTOR asks for information in forms other than numeric probabilities, and then converts this information into numeric probabilities for internal processing [Hart 77]. People describe prior odds to the system by using such terms as "rare" or "common", and describe conditional probabilities using an increased/decreased belief measure.

An earlier (and probably more well known) version of PROSPECTOR used prior odds of events and the probabilities of a piece of evidence conditional on one hypothesis as sources of information. To find the desired conditional probability, these sources of information are combined using Bayes' rule and assumptions of conditional independence between pieces of evidence. This version has problems of assuming too much independence between evidence and some associated ad-hoc and not too successful efforts to get around it (mentioned in [Duda 79]), and also of using inconsistently specified probabilities (mentioned in [Duda 76]). Because of the latter problem, the method indicated by the rules of probability for the updating of the odds of hypotheses in the inference net is replaced by a more or less ad-hoc interpolation scheme [Duda 76, Duda 79] which does not always work well.

A later version of PROSPECTOR (described in the appendix in [Duda 79] by K. Konolige) manages to get around these two problems of the earlier version. This later version treats the given information (conditional and unconditional probabilities supplied by the experts) as constraints on the probability distribution that describes the interdependencies between the events in each group of closely interrelated events. It then estimates the probability distribution for each group by assuming that the various events are as independent as possible, subject to the mentioned constraints. The estimation process involves an information-theoretic measure and a modified linear programming algorithm (the concave Simplex method). In the course of making the estimation, the process checks the given information for consistency, rejecting any set of inconsistent information. However, this newer version does have the problem that it seems to ask people for information in the form of hard to estimate probabilities.

When in use, this newer version gets the desired conditional probability of a hypothesis with the help of an assumption of independence (also made in the earlier version) from the effects of some more remote evidence in the net, when some closer evidence in the net is known. That is, when the net indicates that evidence E is used to find the belief in hypotheses H, and evidence E' is used to find the belief in evidence E occurring, then the assumption is made that when E is known, H is independent of E'. This assumption seems to hold up well in the examples given in the reports on

PROSPECTOR. But it should be noted that no mention is made of any effort to insure that the structuring of the knowledge will maintain this assumption's validity.

2.4 MYCIN

MYCIN [Shortliffe 76, Davis 76] is another computer system designed to aid decision-making. It is designed to give advice on bacterial infection diagnoses. Like PROSPECTOR, MYCIN structures its knowledge using a version of an inference net. Besides being used to show where possible inferences can be made, this structure is used to help generate explanations of judgments the system makes [Davis 76].

Because of the difficulty that people have in estimating probabilities, MYCIN does not have a probabilistic basis. Instead, MYCIN uses a confirmation measure which indicates the increased or decreased belief in one event occurring when another event is known to have occurred. The measure itself has a probabilistic basis. However, this is ignored in favor of certain intuitions as the basis for how MYCIN combines the given measures to obtain the desired measures. For example, given $MB[H|E_1]$ and $MB[H|E_2]$, the increased belief in the hypothesis H given the evidence E_1 and E_2 respectively, MYCIN will assume that $MB[H|E_1 and E_2]$, the increased belief in H given both E_1 and E_2, $= MB[H|E_1] + MB[H|E_2] \times (1 - MB[H|E_1])$. Actually, the methods of combination seem to work well in most cases, but when they do not, there is no means of cleanly overriding them with additional information. Continuing the above example, suppose that an expert knows the value of $MB[H|E_1 and E_2]$, and that this value differs from what MYCIN assumes. The only way that one will be able to correct MYCIN is to do some ad-hoc thing like telling it that $MB[H|E_1 and E_2] = y$, where y (not the expert's value of $MB[H|E_1 and E_2]$) is the value ,if any, that one finds (by trial and error) will make the system behave well. This problem is not uncommon. As mentioned in [Shortliffe 75] (which also describes this measure, the rationale for its use, and the methods of combination), erroneous results often seem to occur when the effects of strongly related pieces of evidence are combined.

A more detailed description of the confirmation measure in MYCIN is given in the chapter outlining my approach for PLY.

2.5 Dempster-Shafer

In [Shafer 76], Glenn Shafer proposes a mathematical theory of evidence. This theory is a generalization of the Bayesian way of treating beliefs (treating beliefs in terms of classical probability).

With Shafer's theory for belief functions, one can express the idea that one has a belief that at least one of a group of n events $X_1,...,X_n$ will occur, but that one does not know how to totally split up that belief of occurrence among the n events. For instance, when those n events are mutually exclusive, Shafer's theory will permit the expression of the following idea: the belief that any one of those events will occur can be greater than the sum of the beliefs of each event occurring ($Bel(X_1 \vee ... \vee X_n) \geq \Sigma_i Bel(X_i)$). With the Bayesian belief system, this idea is not expressible. For this case, this system will enforce the following idea: the belief that any one of those events will occur is equal to the sum of the beliefs of each event occurring ($Bel_p(X_1 \vee ... \vee X_n) = \Sigma_i Bel_p(X_i)$). This enforcement is obvious once one notes that in this case, $P(X_1 \vee ... \vee X_n) = \Sigma_i P(X_i)$.

In a sense, Shafer's belief function, Bel, is a measure of the lower probability of an event's occurrence, since when splitting-up the belief for one of a group of events occurring, some of the belief may not be assigned to any one of the events. This unassigned belief clearly belongs to one or more of the events, but one does not know which one(s).

To complement this idea of Bel being a lower probability, Shafer also defines an upper probability function, P*, which measures the degree to which one fails to doubt an event's occurrence. $P^*(A) = 1 - Bel(\sim A)$.

Weighing against the advantage of being able to express a lack of knowledge about an event's chances of occurring are problems created by using this theory. One of these is that with this theory, one does not have just one probability to deal with for a given event, but two, an upper and a lower one. As mentioned on page 874 of [Barnett 81], it is not clear how one should deal with a bounds on the range, as opposed to a unique value. This is especially true when the bounds are loose. For example, knowing that the probability that a particular event will occur is at least 0.1, and at most 0.9 does not tell one very much about that event beyond that it could occur.

Another problem is that this formalism has an even worse combinatorial explosion problem than classical probabilities. Given n events, one now does not just have to be concerned about the 2^n possible combinations of the events occurring or not occurring, but all the $2^{(2^n)}$ possible subsets of this set of 2^n possible combinations. As an example of this, for the two events A and B, one will have to find Bel of the following 16 possibilities: $\emptyset$, $\sim A \wedge \sim B$, $\sim A \wedge B$, $A \wedge \sim B$, $A \wedge B$, $\sim B$, B, $\sim A$, A, $(\sim A \wedge \sim B) \vee (A \wedge B)$, $(\sim A \wedge B) \vee (A \wedge \sim B)$, $\sim A \vee (A \wedge \sim B)$, $\sim A \vee (A \wedge B)$, $A \vee (\sim A \wedge B)$, $A \vee (\sim A \wedge \sim B)$, and U, where $\emptyset$ is the possibility that no event occurs or does not occur (it gets a Bel of 0), and U is the possibility that some event occurs or does not occur (it gets a Bel of 1).

In addition, a problem of just using this theory to build a plausibility inference system is that, like the classical probability formalism, it does not tell one how to organize the given information so that people checking the system will be able to find out easily what knowledge the system may be missing, or what erroneous information the system might have.

Some of the interest in using Shafer's theory, notably [Barnett 81, Garvey 81], centers around Dempster's rule of combination (orthogonal sum). This rule combines the belief functions of two sources of evidence to get the resulting belief function. Unfortunately, this rule is of limited use. It is meant only for combining independent sources of evidence, and even for these, the rule is justified purely on intuitive grounds.

2.6 Fuzzy Logic

Fuzzy sets and fuzzy logic, as explained in [Gaines 76], became popular after a paper by Zadeh on the subject a few years back. With fuzzy sets, an element can partially belong to a set, as opposed to either totally belonging to a set, or being totally outside a set. Such an idea tries to capture the meaning of such statements as *x is very y*, or *x is only slightly y*. The extent to which an element belongs to a set is measured by its *degree of membership* in that set. The degree of membership is a real number ranging from 0 (element is not in the set at all), to 1 (the element is totally in the set). The boolean operations on fuzzy sets are as follows, given an element x and sets A and B:

Union — The degree that x belongs to either A or B is the maximum of the degree that x belongs to A and the degree that x belongs to B: $(A \cup B)x = \max(Ax, Bx)$.

Intersection — The degree that x belongs to both A and B is the minimum of the degree that x belongs to A and the degree that x belongs to B: $(A \cap B)x = \min(Ax, Bx)$.

Complement — The degree that x doesn't belong to A is 1 minus the degree that x belongs to A: $(\sim A)x = 1 - Ax$.

While the degree of membership of an element in a set in fuzzy set theory looks somewhat like the probability that the element belongs to a classical set, the two notions are different. The former notion says that the element belongs to the set with degree x, while the latter notion says that the element either is in the set, or it is not, and the odds that it is in the set is x. One can see a difference between the two notions in the intersection and union operations, where the fuzzy set operators correspond to the special case for classical sets and probability where either B is a subset of A or vis-versa (complete dependence). From this example, one can see that used straightforwardly, fuzzy sets will not be able to express many conditions expressible with a probabilistic basis. However, despite this lack of generality, the idea of fuzzy sets may prove to be useful when one is trying to break apart a measurement with a continuous range into a series of binary events. An example of this use can be found in [Ishizuka 81].

Fuzzy sets are used as the basis for fuzzy logic, which can be defined as follows: Let F be the set of all false expressions and T be the set of all true expressions. Let the letter "a" represent an arbitrary statement, and $\alpha = (1 - Fa + Ta)/2$ measure the degree of "truthness" of that statement. Similarly, define b and β, and c and γ. Then:

for $b = \sim a$, $\beta = 1 - \alpha$,

for $c = (a \vee b)$, $\gamma = \max(\alpha,\beta)$,

for $c = (a \wedge b)$, $\gamma = \min(\alpha,\beta)$,

for $c = (a \rightarrow b) = (\sim a \vee b)$, $\gamma = \max(1-\alpha,\beta)$,

for $c = (a \equiv b)$, $\gamma = \min(\max(1-\alpha,\beta),\max(1-\beta,\alpha))$.

As mentioned by Gaines, the above definition for implication has a problem. When $c = (a \rightarrow b)$ is always true ($\gamma = 1$), either $\alpha = 0$ (a is always false), or $\beta = 1$ (b is always true). This is in discord with the idea in fuzzy logic that propositions can generally belong to the set of true propositions with any degree between 0 and 1, so many alternatives for the degree of "truthness" of a proposition $a \rightarrow b$ have been proposed. Among them are: (1) if $\alpha \leq \beta$ then 1 else 0; (2) if $\alpha \leq \beta$ then 1 else β; (3) $\min(1,1-\alpha+\beta)$; and (4) $\min(1,\beta/\alpha)$.

People have proposed using fuzzy logic as a basis for "imprecise" reasoning. However, when using fuzzy logic to represent beliefs, the combination functions will often give strange results. For example, take the event "Lisa is skating" (and let it be represented by the letter "A"), and the corresponding degree of belief in it α (as before, $\alpha = (1- FA + TA)/2$). Fuzzy logic then states that the degree of belief in $A \wedge \sim A$ will be $\min(\alpha,1-\alpha)$, which is non-zero for $0 < \alpha < 1$. This means that if we somewhat believe in A, then we will have some belief that Lisa is simultaneously skating and not skating. A similar problem occurs with $A \vee \sim A$, which can result in having some belief that Lisa is neither skating, nor not skating ($\sim[A \vee \sim A]$). Also, all the definitions of implication permit some belief in both A and $C = (A \rightarrow (\sim A))$ at the same time (since all the definitions permit $\alpha>0$ and $\gamma>0$ at the same time, where γ is the belief in C).

2.7 Modal Logic

Modal logic, as described in [Hughes 68], is an extension of predicate calculus. This logic considers a set of worlds in which various statements are true or false, and has the notion of a statement being *impossible*, *possible*, or *necessarily true*. For each world, there is a set of worlds that it can *view* or *access*. In a given world, a statement is considered *impossible* if that statement is false in all the worlds that it can view. Similarly, in a given world, a statement is *possible* or *necessarily true* if that statement is true in at least one or all the worlds that it can view, respectively.

A problem with using modal logic is that there is no idea of measuring how likely an event is to occur. Events either never occur (statement is impossible), will sometimes occur (statement is possible), or will always occur (statement is necessarily true). However, modal logic might provide techniques for "weeding out" events that are either impossible or necessarily true, and then leaving it to another system to determine how likely the possible events are.

3. The PLY Approach: An Overview

The following is an overview of the approach that I will be taking to develop PLY. In my thesis, I will develop this approach in detail, and show how one would use PLY to aid in planning. To the extent that time permits, I will also start to implement PLY.

3.1 Events and Probability Distributions

This system will deal with possible events by trying to determine the chances of, or beliefs in, the occurrence of those events. A possible event here is anything that is expressible as an proposition in English, logic, etc. Examples of possible events include the propositions "Joe is playing with a ball", and "Orson is a person".

Given n possible events, there are 2^n possible combinations of those events occurring and not occurring (each possible event can occur or not occur). If one assigns a probability of occurrence to each of these combinations, these probabilities will collectively form a joint probability distribution for the events concerned. For this distribution to be a valid probability distribution, it must satisfy two requirements: one is that the probability of every combination is a real number between 0 and 1 (inclusive), and the other is that the sum of the probabilities of all 2^n combinations equals 1.

Once this distribution is found, one can find the current chances of any event occurring by using the distribution to find its probability conditional on the events that one knows have occurred. To find this distribution, one must get information from experts and/or statistics on the events and their relationships.

3.2 Correlation and Confirmation

As mentioned in the introduction, people will provide much of the information for finding the joint probability distribution, and they generally have a hard time trying to estimate probabilities. So instead of asking people ('experts' and users) to only give information in the form of conditional or unconditional probabilities, PLY should also let them give information in the form of correlations and/or the confirmation measure used in MYCIN. Both measures take a value from -1 to 1, with 0 implying that the events involved are independent.

One can define the correlation between events by using a version of the correlation coefficient, which is a measure of the normalized correlation between two random variables. Let each event or combination of events have a corresponding Bernoulli random variable. This random variable will take on the value of 1 when the event or combination occurs, and 0 when it does not. Then define the correlation of two events (or combinations), A and B, as the correlation coefficient of their respective random variables (r.v.'s):

$$(m_{AB} - m_A m_B)/(\sigma_A \sigma_B) =$$

$$[P(A,B) - P(A)P(B)]/[P(A)(1-P(A))P(B)(1-P(B))]^{0.5},$$

where m_{AB} is the expected value of the r.v. for the event that both A and B occur, m_X is the expected value of event X's r.v., and σ_X is the variance of event X's r.v.. This will measure in a sense how non-independent two events are. The more positive the measure for the two events (or groups of events), the more the two tend to occur together when they do occur, and the more negative the measure for two events, the more they tend to not occur together when they do occur. A correlation of zero means that the occurrence of one event does not seem to tell anything about the occurrence of the other event.

The confirmation measure in MYCIN [Shortliffe 75] of how much event A's occurrence is confirmed by event B's occurrence is a bit more complicated, and is defined as:

```
CF[A,B] = if $P(A) = 1$ then 1
          else if $P(A) = 0$ then -1
          else if $P(A|B) \geq P(A)$ then $[P(A|B)-P(A)]/[1-P(A)]$
          else $[P(A|B)-P(A)]/P(A)$.
```

The more positive CF[A,B] is, the more that knowing event B occurred will imply that event A also occurred, while the more negative it is, the more that knowing event B occurred will imply that event A had not occurred. When CF[A,B] = 0, event B's occurrence will not alter the chances of A occurring.

These two measures should be easier for people to estimate than probabilities, since in thinking about how two events are related, people tend to think in terms of how much knowing about one event's occurrence changes the chances of another event occurring (that is, how strongly the occurrences of the two events seem to be connected or mutually exclusive), as opposed to the numeric odds of two events occurring at the same time (a joint probability), or the numeric odds of one event occurring when a second event occurs (a conditional probability). For example, it is easier to estimate how much knowing that a person has a high grade point average tends to increase or decrease one's belief that that person will be a good claims examiner than to estimate the corresponding numeric odds.

3.3 Information as Constraints

One way of looking at the information given to the system by experts and/or statistics on combinations of events is that they are constraints on the possible probability distributions that these combinations of events could have. For example, if one considers the two events A and B, and is told that P(A) = 0.4, then one knows that the joint distribution for A and B, $p_{AB}(x,y)$, cannot be just any joint distribution for two events, but must be one in which P(A) = P(A,~B) + P(A,B) = 0.4. This says that, within the distribution, the probabilities of two particular combinations, P(A,~B) and P(A,B), have to sum to 0.4. Or if one is given P(B|A) > 0.1, then the joint distribution has to satisfy P(B|A) = P(A,B)/P(A) = P(A,B)/[P(A,~B) + P(A,B)] > 0.1.

Although the resulting constraint relationships are even more complicated than those above, correlation and confirmation information can both also be thought of as constraints on the probability distribution. For instance, for the example above, correlation(A,B) can be expressed in terms of the combination probabilities as $[P(A,B)\text{-}XY]/[X(1\text{-}X)Y(1\text{-}Y)]^{0.5}$, where X = P(A,~B) + P(A,B), and Y = P(A,B) + P(~A,B).

This view of using the given information as constraints on the distribution has been taken by Kurt Konolige in developing an advanced version of PROSPECTOR [Duda 79], a mineral consultation program. It may be contrasted with approaches used in other systems, such as MYCIN [Shortliffe 76], and an earlier version of PROSPECTOR [Duda 79], both of which automatically assume certain things about the distribution, and so will ignore any information that conflicts with those assumptions when it answers queries about that distribution. Actually, these systems can be made to take this information into account. However, to do so usually involves doing some *ad hoc* thing like telling the system one value to try to get it to act as if it were another value. Having to do such things tends to make the acquisition and inspection of knowledge for the system much harder, since the relationship between the system's knowledge and how the system will behave becomes more obscure. In other words, these systems cannot cleanly handle many perfectly valid situations. An example of this problem in MYCIN was given in an earlier chapter surveying existing approaches and systems. An example of this in PROSPECTOR is discussed in [Duda 79] in the appendix by Konolige. It involves the combination of evidence for the presence of a Komatiite rock suite (KRS), and shows the problems of trying to combine non-conditionally independent evidence in the older version.

3.4 Assumptions and Estimation

After the constraints are applied to a distribution, one can get three possible results:

1. The constraints are inconsistent. No probability distribution exists which will satisfy all the constraints simultaneously. The constraints describe an impossible situation.

2. Exactly one probability distribution will satisfy all the constraints. That is, every combination gets a unique probability value.

3. More than one probability distribution will satisfy all the constraints. The combinations can be assigned more than one set of probability values which will satisfy the constraints.

How to treat the first two cases is fairly clear. In the first case, one has to find and repair the errors in the constraints, so that one ends up with the second or third case. In the second case, one has found the distribution that one is after.

If more than one probability distribution will satisfy the constraints (case 3), one has two options:

1. Whenever PLY encounters something that can take on a range of values, it can report the upper and lower bound on that range. In this case, something like the constraint system used in the computer vision system of Brooks [Brooks 81] can be used to determine if the constraints are consistent, and if so, determine the bounds on the range for the values concerned.

2. PLY can try to find best estimates for the probabilities of all combinations and use them as "the" distribution.

The first option gives an accurate picture of the implications of the given information. However, as mentioned in the section on Shafer's theory in the survey chapter, it is not clear how effectively one will be able to work with bounds on the range, especially when they are loose. To get answers with tighter bounds, one will have to collect more and more information on the events, even information which just says that two or more particular events have nothing to do with each other (are independent).

The second option has the advantage of providing unique values, rather than just upper and lower bounds. However, to find a best estimate, one will have to come up with some criteria or measure for a best estimate, and this means making assumptions. Having assumptions is not all bad, however. If one knows what the system will assume about missing information, then one can just not bother to explicitly give information that confirms what the system will assume anyhow. Besides potentially not having to bother the experts and statisticians as much, this has the advantage that systems to manipulate constraints will run faster when fewer explicit constraints are present.

One choice for a best-estimate probability distribution is the one in which the events are as independent as possible (knowing whether one event occurs or not tells as little about the possibilities of the other events as possible), given the constraints (information) present. In other words, for two events A and B, the distribution should have P(A|B) be as close to P(A) as the constraints will permit, and similarly for P(B|A) and P(B). This method makes the assumption that if the expert(s) consulted did not bother to mention a relationship between two possible events (that yields a constraint), then knowledge about one of the possible events should not affect knowledge about the other possible event to the extent allowed by other relations given. In other words, this method assumes that experts will tend to express strong relationships between events (the ones that are most non-independent) and tend not to bother with weaker ones. Intuitively, this assumption is a good one to make. For example, when asked about what pieces of evidence will tend to increase one's belief that a particular person would make a good final assembly inspector, one would think that the person questioned would very likely to mention such traits as patience and meticulousness, and would very rarely mention a trait such as liking to eat liver. It is fairly evident that liking liver has little to do with being a good inspector, but unless PLY either makes this assumption or is explicitly told about this very weak (or non-existent) relationship, it cannot use this belief to constrain the probability distribution.

Another choice for a best-estimate probability distribution is the distribution which satisfies the constraints and which favors as little as possible the occurrence of one combination of events over any other combination of events. That is, the distribution to be used is that which satisfies the constraints and in which the combination of events which will occur is the most unpredictable (random). The justification here is that, to the extent that one's information does not tell one the odds of the combinations of events occurring, one should try not to favor the occurrence of any one combination over any other combination.

It turns out that both of the choices discussed above can be shown to be equivalent to picking that distribution satisfying the given constraints which minimizes the value of the information-theoretic function

$I_p \equiv \log 2^n - H_s$

$= \log 2^n - (- \Sigma P_j \log P_j)$

= the maximum possible information content in a distribution

- the average information given by an observance of which possibility occurs,

where n is the number of events that one is concerned with, and P_j is the probability of the jth combination of events. This function measures the information content of a probability distribution in an information-theoretic sense, and is mentioned in [Lewis 59]. In the advanced version of PROSPECTOR referred to earlier, Konolige used this function for estimating the best probability distribution, given a set of constraints. Note that this method of making assumptions has the advantage that the assumptions can always be overridden where they are not tenable. One just has to add and/or replace constraints to achieve the desired corrections.

Finding the probability distribution that minimizes I_p for a given set of constraints is a constrained optimization problem. To solve such problems, one can use various programming algorithms. For certain kinds of constraints, the concave Simplex method [Wagner 69] or a method developed in [Brown 59] may be used; otherwise, a more general non-linear programming method may be needed.

It should be noted that the two options of giving bounds or finding a 'best' estimate are not mutually exclusive. One could give both a range limit and a best estimate, with the former providing a sense of how much the estimate could be off.

3.5 Structuring Knowledge

So far, everything that has been said implies that, to build a consultation system, one should first obtain expert and statistical information to determine the relevant events and their interrelationships, and then find the best estimate joint distribution and probabilistic range limits for the events and their combinations. However there is a very basic problem with this approach. Given *n* events, there are 2^n combinations of events for which one has to find the probabilities. So, for any non-trivial *n*, one will have to be concerned about an enormous number of combinations. For instance, if there are 17 events, then there are $2^n = 2^{17} =$ 129,536 combinations to consider, which is an enormous number.

One way of getting around this problem is to structure the given knowledge in a form similar to the organization of possible beliefs and their relationships in Doyle's model of deliberation and introspection [Doyle 80a]. This structuring will reflect the links between hypotheses and their possible evidence. These links can reflect one of several types of relationships, including such common ones as:

1. classification or decomposition: A given event (a hypothesis) may be decomposed into a group of events (possible evidence). An example is that having a good background for the computer science written exam at MIT (CSWE) can be decomposed into having a good background in each of the four parts of the exam: programming languages, algorithms and complexity, computer architecture, and artificial intelligence.

2. causality: The (lack of) occurrence of a given hypothetical event may be caused by (earlier) occurrences of other event(s) (possible evidence). An example is that studying a subject tends to cause one to be knowledgeable about that subject. Causality has been the focus of much work in artificial intelligence, examples being [Patil 81], and CASNET [Kulikowski 73, Szolovits 78].

3. tests or signs: The occurrence or non-occurrence of a given hypothesis tends to be revealed by the (possibly later) occurrence or non-occurrence of some other events (possible evidence). One example is that having a fever or swelling around a cut tends to indicate that bacteria had infected that cut. Another example is that watching a piece of metal catch on fire in water is evidence that the metal is sodium or potassium.

The structuring should be such that the events are arranged in "layers", where events in each layer provide reasons (evidence) to believe or disbelieve in events in the layer immediately above it, and will affect the beliefs about of events in layers farther above it only because of their influence on the layer(s) in-between. An example is the following. Suppose that event A is taking the basic course in artificial intelligence, event B is having a good background in artificial intelligence, event C is having a good background for the CSWE, and the goal is to find the belief in event C. A's occurrence will increase the belief that B and C will occur. B's occurrence will also increase the belief that C will occur, and furthermore, A's occurrence tends to increase the belief in C's occurrence only because A's occurrence tends to increase the chances of B occurring. Since the goal is to find the belief in C's occurrence, C should be at the highest layer, and B should be in a layer below. Finally, because A seems to effect C only because of A's effect on B, A should be in a layer below B. This layering permits a certain form of conditional independence to be assumed, which allows the joint probability distribution for all the events to be approximated from the distributions for events in adjacent layers. As mentioned before, this assumption was used in the same way in PROSPECTOR [Duda 76, Duda 79] to simplify the updating of probabilities of hypotheses in PROSPECTOR's net of information. Such a structuring of PLY's knowledge will also make explicit which groups of events within a layer are totally independent of each other, so the joint probability distribution for all the events in these groups can also be approximated by simple combinations of smaller probability distributions. This ability to use many smaller joint distributions in place of one larger one will result in storage savings because, as we have seen, the size of a distribution is exponential in the number of events in that distribution.

PLY's structuring of knowledge also has advantages for acquisition of information used to find distributions, for making changes in PLY's knowledge, and for PLY-generated explanations of its decisions. The layering aids the acquisition of relevant information by helping to identify which relationships PLY should try to get good estimates for (the ones between the same and adjacent layers of events) and which ones PLY can automatically take care of (the ones between events separated by one or more layers). In effect, the structuring decomposes the problem of finding the relationships between events to that of finding the relationships between events in the same and adjacent layers. Of the relationships present, these tend to be the easiest for people to recall and to estimate the strength of[1] .

The structuring of the knowledge will also help the people giving knowledge to PLY see what events and/or interrelationships of importance might be missing, extraneous, or otherwise in need of alteration. For example, it is easier to see if a certain relationship is missing from the system when the knowledge is organized, than when it is arranged in a random fashion. This advantage has been noted for the PROSPECTOR system [Hart 77, Duda 79] in a survey by Kulikowski [Kulikowski 79]. Similarly, this structuring will help people to make temporary changes in the system. Such temporary changes may be desired to test hypothetical situations, or to temporarily use a set of slightly different criteria. The latter can easily arise, for different experts and users can have different opinions on how important something is, or even what is important in the first place.

Besides making it easier for people to figure out what temporary changes they might want to make, this structuring also makes it easier for the temporary changes to be implemented on the system. When something changes, one will not have to recompute the entire distribution for all the events, but only those much smaller distributions which are affected. This means that the system will be able to adjust to these temporary changes much more quickly. It also means that the system will be able to store many more versions of its knowledge base (each reflecting a few small change(s) from the rest), since it can store new versions just by storing the few changes from older versions. This idea of sharing between versions has previously used in [Stallman 76], and in connection with *truth maintenance systems* [Doyle 79, McAllester 80, Doyle 80b] and the planning graphs of the Programming Technology Division at the MIT Laboratory for Computer Science.

Past builders of knowledge-based systems have shown that such structuring helps in making acceptable explanations of system actions to users. This is because such structuring helps identify the intermediate states in the reasoning process, so that when asked to give the factors used in arriving at an answer, the system can give some chains of reasons by noting the intermediate factors (events) involved, as opposed to just reciting the favorable evidence the user gave. For example, having the system say that a certain person would be a good claims examiner because that person likes to make small wood carvings would leave one to wonder how the system arrived at such a decision. However, if the system said that a certain person would be a good claims examiner because it has evidence that the person is patient and meticulous, and this evidence is that the person likes to

[1] The advantage of decomposition is fairly widely accepted: witness the fairly wide push in programming methodology for the concept of functional decomposition. [Pearl 82] gives a short discussion of the rationale for decomposition in problem-solving, and cites some research on the matter.

make small wood carvings, then the explanation would much more acceptable. A good example of using structured knowledge to produce explanations can be found in the MYCIN project, and is described in [Davis 76].

3.6 Summary

The PLY approach borrows many ideas from the works of others in artificial intelligence. Among these are:

1. The probabilistic approach taken by Konolige.
2. The confirmation measure used in MYCIN.
3. The bounds calculation system of Brooks.
4. The structuring of knowledge found in MYCIN and PROSPECTOR.
5. Doyle's model of deliberation.
6. The sharing of information between contexts, as used in truth maintenance systems and planning graphs.

My contribution will be to try to integrate these ideas in developing PLY, using Konolige's approach as the basis. In addition, I will present new arguments for using Konolige's approach and introduce the use of an event correlation measure.

4. Bibliography

[Barnett 81]

Barnett, Jeffrey.
Computational Methods for a Mathematical Theory of Evidence.
In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 868-875. The International Joint Conferences on Artificial Intelligence, August, 1981.

[Brooks 81]

Brooks, Rodney A.
Symbolic Reasoning Among 3-D Models and 2-D Images.
Artificial Intelligence 17:285-348, 1981.

[Brown 59]

Brown, David.
A Note on Approximations to Discrete Proabability Distributions.
Information and Control 2:386-392, 1959.

[Davis 76]

Davis, R.
Applications of meta level knowledge to the construction, maintainance and use of large knowledge bases.
AIM 283, Stanford AI Lab, 1976.

[Doyle 79]

Doyle, Jon.
A truth maintaince system.
Artificial Intelligence 12:231-272, 1979.

[Doyle 80a]

Doyle, Jon.
A Model for Deliberation, Action, and Introspection.
AI TR 581, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., May, 1980.

[Doyle 80b]

Doyle, Jon and London, Philip.
A Selected Descriptor-Indexed Bibliography to the Literaure on Belief Revision.
A.I. Memo 568, MIT Artificial Intelligence Lab, Feb, 1980.

[Duda 76]

Duda, R. O., Hart, P. E., and Nilsson, N. J.
Subjective Bayesian methods for rule-based inference systems.
In *AFIPS Conference Proceedings, Vol. 45*, pages 1075-1082. National Computer Conference, 1976.

[Duda 79]

Duda, R., Hart, P., Konolige, K., Reboh, R.
A Computer-Based Consultant for Mineral Exploration.
Final Report, SRI Project 6415, SRI International, Menlo Park, CA, September, 1979.

[Gaines 76]
Gaines, B. R.
Foundations of Fuzzy Reasoning.
International Journal of Man-Machine Studies 8(6):623-668, November, 1976.

[Garvey 81]
Garvey, Thomas, Lowerance, John, and Fischler, Martin.
An Inference Technique for Integrating Knowledge From Disparate Sources.
In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 319-325. The International Joint Conferences on Artificial Intelligence, August, 1981.

[Hart 77]
Hart, P.E., and Duda, R.O.
PROSPECTOR - A Computer-Based Consultation System for Mineral Exploration.
TN 155, SRI International, Menlo Park, California, October, 1977.

[Hughes 68]
Hughes, G., Cresswell, M.
An Introduction to Modal Logic.
Methuen and Co Ltd, London, 1968.

[Ishizuka 81]
Ishizuka, M., Fu, K., Yao, J.
Inexact Inference for Rule-Based Damage Assessment of Existing Structures.
In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 837-842. The International Joint Conferences on Artificial Intelligence, August, 1981.

[Kahneman 82]
Kahneman, Danial, and Tversky, Amos.
The Psychology of Preferences.
Scientific American 246(1):160-173, January, 1982.

[Kulikowski 73]
Kulikowski, C., Safir, A., Weiss, S.
A Representation of Medical Knowledge for Problem Solving: Application to a Model of Glaucoma.
CBM-TR 21, Dept. of Computer Science, Rutgers University, New Brunswick, N.J., July, 1973.

[Kulikowski 79]
Kulikowski, C.
Artificial Intelligence Methods and Systems for Medical Consultation.
CBM-TR 101, Dept. of Computer Science, Rutgers University, New Brunswick, N.J., Sept, 1979.

[Lewis 59]
Lewis, P. M. II.
Approximating Probability Distributions to Reduce Storage Requirements.
Information and Control 2:214-225, 1959.

[McAllester 80]

McAllester, David.
An Outlook on Truth Maintenance.
AI Memo 551, MIT Artificial Intelligence Lab, Aug, 1980.

[Patil 81]

Patil, R., Szolovits, P., Schwartz, W.
Causal Understanding of Patient Illness in Medical Diagnosis.
In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 893-899. The International Joint Conferences on Artificial Intelligence, August, 1981.

[Pearl 82]

Pearl, J., Leal, A., Saleh, J.
GODDESS: A Goal-Directed Decision Structuring System.
IEEE Trans. on Pattern Analysis and Machine Intelligence PAMI-4(3):250-262, May, 1982.

[Pednault 81]

Pednault, E., Zucker, S., Muresan, L.
On the Independence Assumption Underlying Subjective Bayesian Updating.
Artificial Intelligence 16(2):213-222, May, 1981.

[Raiffa 68]

Raiffa, Howard.
Decision Analysis, Introductory Lectures on Choices Under Uncertainty.
Addison-Wesley Publishing Co., Reading, MA, 1968.

[Shafer 76]

Shafer, G.
A Mathematical Theory of Evidence.
Princeton University Press, Princeton, 1976.

[Shortliffe 75]

Shortliffe, E. H., and Buchanan, B. G.
A Model of inexact reasoning in medicine.
Mathematical Biosciences 23:351-379, 1975.

[Shortliffe 76]

Shortliffe, Edward.
Computer-Based Medical Consultations: MYCIN.
American Elsevier Publishing Co., Inc., N.Y., 1976.

[Shortliffe 79]

Shortliffe, E., Buchanan, B., Feigenbaum, E.
Knowledge Engineering for Medical Decision Making: A Review of Computer-Based Clinical Decision Aids.
STAN-CS 79-723, Stanford U., February, 1979.

[Stallman 76]

Stallman, R., and Sussman, G.
Forward Reasoning and Dependency-Directed Backtracking In a System for Computer-Aided Circuit Analysis.
Memo 380, MIT Artificial Intelligence Laboratory, September, 1976.

[Szolovits 78]

Szolovits, Peter, and Paulker, Stephen.
Categorical and Probabilistic Reasoning in Medical Diagnosis.
Artificial Intelligence 11:115-144, 1978.

[Szolovits ??]

Szolovits, Peter.
Remarks on Scoring.
draft of April 1976 at the MIT Lab for Computer Science

[Wagner 69]

Wagner, H.
Principles of Operations Research.
Prentice-Hall, Inc., 1969.