Case Based Predictions: Introduction*

Itzhak Gilboa and David Schmeidler

December 2010

1 Background

There are several approaches to formal modeling of uncertainty, knowledge, and belief. They differ in the way they represent what is known and what is not known, the formal entities that capture beliefs, the manner in which these beliefs are updated in the face of new evidence, and so forth. For example, classical statistical inference describes knowledge by a family of distributions, where the reasoner is assumed to know that the process is governed by one of these distributions, but she does not know which one. Evidence is modeled as realizations of random variables, and beliefs are updated according to classical techniques such as maximum likelihood estimation, the construction of confidence sets etc. By contrast, the Bayesian approach represents knowledge as a set of states of the world, such that the reasoner knows the set but does not know which particular state in it obtains. Beliefs are represented by a prior probability measure over the state space, while evidence is modeled as events, namely subsets of states, such that belief revision consists in Bayesian updating of the prior probability to a posterior. A rather different approach considers rules to be the primary objects of knowledge, while evidence is modeled as particular instances in which rules may or may not

*Introduction to "Case-Based Prediction", a volume to be published by World Scientific Publishers; Economic Theory Series, edited by Eric Maskin.

hold. The rules represent beliefs, and they are updated, in light of new information, according to belief revision methods (see Alchourron, Gardenfors, and Makinson, 1985 and Gardenfors, 1992). Another approach assumes that the object of knowledge are particular cases, or observations, where beliefs are indirectly expressed by the similarity one finds between different cases (Schank, 1986, Riesbeck and Schank, 1989). And one may also describe knowledge and belief by neural nets, fuzzy sets, and other methods.

Given the state of art in the behavioral and social sciences on the one hand, and the limitation of the different methods on the other hand, it stands to reason that no single method would dominate all the others. Rather, one would typically expect that each method would have some applications to which it is best suited, and others where it may be inconvenient or awkward. Indeed, the literature in statistics, machine learning, artificial intelligence, and engineering is rather pluralistic. Even economists, when conducting research, use several methods. However, when modeling a rational agent, the modern nickname of homo economicus, the latter is restricted to be Bayesian. This is partly a result of the success of game theory. The concept of strategic (Nash) equilibrium replaced and extended that of price equilibrium, and greatly enhanced the ability of economists to analyze interactive situations. The equilibrium as defined by Nash requires mixed strategies, that is, beliefs that are quantified probabilistically. Moreover, the applicability of games to economics was further extended with Harsany's modeling of incomplete information in a Bayesian way (introducing the concept of Bayesian equilibrium). Modeling economic agents as non-Bayesians may cast doubts on the usefulness of game theory and the validity of its economic conclusions.

Another source of support for the assumption of the Bayesian agent is its simplicity and its impressive axiomatization. (For the latter see Ramsey, 1931, de Finetti, 1931, 1937, von Neumann and Morgenstern, 1944, and Savage, 1954.[1]) In the Bayesian approach the objects of knowledge and be-

[1] We discuss the meaning and goals of axiomatizations below. See also Gilboa (2009) for

lief, namely events, also model the evidence one may obtain. Further, this approach has only one type of updating, that is, Bayesian updating. In comparison to the variety of classical statistical inference techniques, or the various theories of belief revision in rule-based paradigms, Bayesian updating of a prior to a posterior shines through as a simple, almost inevitable updating procedure that suffices for all purposes. Moreover, the Bayesian approach is tightly linked to a decision-making procedure, i.e., expected utility maximization, and the two can be jointly derived from very elegant axioms.

However, despite its elegance and generality, its axiomatic foundations and breadth of applications, the Bayesian approach has been criticized on several grounds. First, it has long been claimed that there are types of uncertainty that cannot be quantified by probabilities. (Knight, 1921, Keynes, 1921, Ellsberg, 1961, Schmeidler, 1989.) Following Knight, the literature distinguishes between situations of "risk", with known probabilities, and "uncertainty", where probabilities are not known. The Bayesian approach holds that any uncertainty can be reduced to risk, employing subjective probabilities. Yet, there is ample evidence that people often behave under uncertainty differently than under risk, and many authors also justify such behavior as rational. Specifically, it has been argued that the Bayesian approach is well suited to describe knowledge, but that it is poor at describing ignorance.

Second, while the existence of subjective probabilities can be justified by seemingly compelling axioms on behavior, the Bayesian approach says little about the origins of such probabilities. The axiomatic derivations suggest that one should have such beliefs, but not what they should be or where they should be derived from. Indeed, when one can provide a good account of the emergence of probabilistic beliefs, these beliefs tend to be objective, because there are good reasons to adopt them and not others. It is precisely when one finds little to say about the origin of beliefs that one needs to resort

general methodological discussions as well as descriptions of these classical contributions and their critique.

to subjectivity.

Third, in the Bayesian approach beliefs are defined on states or events. By contrast, in economics data are usually collected and presented as lists of observations or cases. If economic agents derive beliefs from data, it may be more intuitive to use models that formally distinguish between data and beliefs, or between observations and theories.

We find that these weaknesses of the Bayesian approach are related. If we have a model of how beliefs are generated, we would know when beliefs would take the form of probabilities, and also when one might seek other models of beliefs. Thus, we find that it would be fruitful to find out which probabilities are chosen by an individual when beliefs are probabilistic, and also which other models of beliefs can be useful when probabilities are too restrictive.

2 Alternative Theories

2.1 Uncertainty

In the early 1980s, the second author developed a theory of decision making under uncertainty that could accommodate non-quantifiable uncertainty, that is "uncertainty" in the language of Knight (1921), "true uncertainty" as referred to by Keynes (1921), or "ambiguity" if one adopts the term suggested by Ellsberg (1961). The theory (Schmeidler, 1986, 1989) involved "probabilities" that were not necessarily additive, with respect to which one can compute expectation using a notion of integration due to Choquet (1953-4). This was the first axiomatically-based general-purpose theory of decision making under uncertainty that generalized the Bayesian approach, and that could smoothly span the entire spectrum between the Bayesian model and a model of complete ignorance.

We later developed the theory of maxmin expected utility (Gilboa and Schmeidler, 1989), holding that a decision maker's beliefs are given by a set

of probability measures ("multiple priors"), and decisions are being made so as to maximize the worst-case expected utility (when probabilities are taken from the prescribed set). This theory is also axiomatically-based, flexible enough to model any decision problem that the Bayesian approach can model, and allows for a continuum of degrees of uncertainty. At the same time, Bewley (2002) developed a theory that also relies on a set of probabilities, modeling a partial order that is defined by unanimity: preference for one option over another only occurs where the former has a higher expected utility than the latter according to each and every probability in the set. In the years that followed, additional models have been suggested, among them are the "smooth" model (Klibanoff, Marinacci, Mukerji, 2005, Nau, 2006, Seo, 2008) and the model of "variational preferences" (Maccheroni, Mukerji, and Rustichini, 2006). For a survey, see Gilboa and Marinacci (2010).

It should be stressed that this line of research has not tackled the question of belief formation. The models mentioned above suggest various generalizations of the Bayesian approach. They are axiomatically based in the sense that one has characterizations of the modes of decision making that are compatible with the formal model. Hence, one can in principle tell whether a particular pattern of choices is compatible with each of these models. But they remain silent on the question of the origin and generation of beliefs, whether probabilistic or not.

2.2 CBDT

In the 1990s we developed a theory of case-based decisions (CBDT, Gilboa and Schmeidler, 1995, 2001). The motivation was to take a fresh look at decision making under uncertainty, and focus on intuitive cognitive processes. Specifically, we sought to develop a formal, axiomatically-based theory that relies on the assessment of past cases rather than of future events. In doing so, we took an extreme approach, and veered away from any notion of belief. Thus, the agents in CBDT do not explicitly have beliefs about future

paths that may unfold should they take various actions. Instead, they are postulated to choose actions that did well in similar cases in the past.

CBDT has several versions. In particular, one can define the notion of "similarity" over decision problems alone, over problem-act pairs, or over entire cases (where a case consists of a problem, an act, as well as a result). It has two versions, one using a summation over cases, and the other – averages; and it can be augmented by a theory of the behavior of an aspiration level that naturally pops up in the analysis.

While the more advanced versions of CBDT are general enough to embed Bayesian expected utility, the fundamental nature of the exercise was not a generalization of the classical theory. Rather, the CBDT was focusing on a particular mode of behavior, awaiting a more general theory that would be able to elegantly encompass both Bayesian, probability-based decision making, and analogical, case-based decisions.

2.3 The present project

The present collection includes papers that deal, for the most part, with case-based predictions. The basic motivation is to use the conceptual basis and mathematical techniques that were developed for CBDT and to apply them to the question of belief formation. We wish to study situations in which beliefs explicitly exist, and might even be given as probability distributions, but to focus on the cases that gave rise to these beliefs. This project is based on the premise that studying the relationship between observations and beliefs may simultaneously shed light on the two questions discussed above: when do beliefs take a particular form, most notably, probabilities, and, when they do, which beliefs emerge from a given database of observations.

Most of the papers collected here do not deal with decisions at all. Rather, they discuss predictions as the outcome of the model. At times, there is an implicit assumption that these predictions are used according to a certain decision theory; specifically, probabilities are assumed to be used for expected

utility maximization. Yet, the formal models ignore decisions. In this sense, this project is closer to statistics than to standard decision theory. On the other hand, it is closer to decision theory in terms of method, in particular in its focus on axiomatic derivations. Before we describe the project in more detail, a few words on the axiomatic approach are in order.

3 The Axiomatic Approach

3.1 Axiomatizations

Axioms generally refer to propositions that are accepted without proof, and from which other propositions are derived. Typically, the axioms are supposed to be simpler and more intuitive than their implications: the rhetorical use of axioms starts with propositions that are accepted and proceeds to those propositions that the listener is supposed to be convinced of. In mathematics and related fields, an "axiomatic system" refers to a set of conditions that captures the essence of a particular structure. Thus, the axioms abstract away from details, generalize the structure and show what are its essential building blocks that are necessary to certain conclusions of interest.

The use of "axiomatizations" in economics refers to conditions on observable data that imply, or even perfectly characterize a certain theory. For example, von-Neumann and Morgenstern's (vNM) axioms on decision under risk are equivalent to the existence of a utility function whose expectation is maximized. This usage of the term "axioms" has much in common with the previous usages. First, vNM's axioms such as transitivity or independence are supposed to be simpler and more intuitive than the explicit representation of expected utility theory. Such an axiomatization is useful for rhetorical purposes, in line with the use of axioms in logic: someone who accepts the axioms is compelled to accept their conclusions. This is obviously useful for normative purposes, because the nature of a normative exercise is precisely this: to convince the listener that a certain mode of behavior (such as ex-

pected utility maximization) is to be preferred. Moreover, the rhetoric of axioms is also useful for descriptive purposes: a scientist who argues that people tend to be expected utility maximizers will be more convincing if she uses simple, acceptable axioms than if she were to use a more complicated theory, despite the fact that the two may be equivalent.

Second, an axiomatization of expected utility theory such as vNM's also serves the purpose of dissection: the axioms state precisely what is assumed by the theory. This simplifies the task of testing whether the theory is correct, and it paves the way to refining or generalizing it in case it isn't. Indeed, a violation of expected utility theory can be analyzed in light of the axioms. One can find which axiom fails, and perhaps indicate why and how it can be relaxed.

However, axiomatizations in economics are assumed to satisfy an additional condition, which was inspired by the thinking of the logical positivists in the philosophy of science (see Carnap, 1922): the axioms are supposed to be stated in terms of observable data. The fact that such axioms imply, or better still, characterize, a theory stated in terms of theoretical concepts renders the latter meaningful. Thus, vNM's axiomatization is viewed as endowing the term "utility" with scientific meaning, showing how it can relate to observations. Relatedly, a theory that does not satisfy axioms stated in observable terms should be suspected of being unrefutable, and therefore non-scientific according to Popper (1934). Axiomatizations therefore guarantee that the game we play has a scientific flavor and that, at least in principle, theories can be tested, and competing theories are guaranteed to have different observational implications.

Despite the fact that logical positivism and Popper's notion of refutation have been seriously challenged within the philosophy of science, we find that they still serve as useful guidelines for economics. It is generally a good idea to ask ourselves questions such as, "What does this term mean precisely?" "Under which conditions will we admit that our theory is false?" The logical

positivist ideal, and axiomatizations in particular, suggest a healthy exercise regime that helps us guarantee that we have satisfactory answers to such questions.

3.2 Axiomatizing predictions

The axiomatizations presented in this volume differ from classical axiomatizations in decision theory in that the observable data they refer to are not choices or preferences but predictions or beliefs. These data are a step removed from economic activity, which involves decisions such as buying and selling, and this fact is often viewed as a disadvantage, at least as far as economics is concerned. On the other hand, these data are independent of the particular decision model one has in mind, and they are closer to the prediction choices made by statisticians, forecasters, or classifiers. Importantly, the axiomatizations relate theoretical concepts to some data that can be observed, and they allow one to ask which pairs of theories are different in content and which pairs may seem to be different while they are, in fact, equivalent.

The prediction problems we are interested in are close to statistical inference. In statistics as well as in machine learning, one deals with questions that are fundamentally very similar to ours: how should one learn from a database of observations? Indeed, we will mention some techniques that are well-known in these fields, such as maximum likelihood estimation or kernel classification. However, the statistical literature does not typically address questions of axiomatization, and focuses instead on asymptotic behavior. Thus, there is a very rich theory about the prediction models that would guarantee satisfactory long-run behavior, but relatively little about the behavior of such models in small databases. Since there are many important situations in which one is asked to make predictions (or to take decisions) despite the paucity of truly relevant past observations, we find this problem to be of interest. We do not believe that one can expect theoretical argu-

ments to provide clear-cut predictions out of thin air. But we hope that the axiomatic approach can at least guarantee that the totality of predictions one offers are coherent in a well-defined sense.

3.3 Statistics and psychology

Our analysis can be interpreted both descriptively and normatively. That is, one may ask which patterns of learning from data are likely to be observed by real people who make predictions in economic contexts; and one may also ask which patterns are desirable, or sensible, and which can serve as ideals of rational reasoning. There are many applications in which the two interpretations can coexist. For example, after observing 100 tosses of a tack, of which 70 resulted in the pin pointing up, it makes sense to predict the same outcome in the next toss. Moreover, this is what most people would do. Thus, the normative recommendation and the descriptive theory coincide in this case. However, there are situations in which the normative differs from the descriptive. In the above example, the "Gambler's fallacy" phenomenon (Tversky and Kahneman, 1974) shows that people might predict that a sequence ("run") of several Heads will be followed by a Tail. In such situations, we typically choose axioms according to their normative interpretation. In other words, when a modeling choice is to be made, we tend to find ourselves closer to statistics than to psychology. One reason for this preference is that axioms are readily applicable to the normative question of choosing among learning methods; by contrast, it is less obvious that one may gain much by axiomatizing the way people actually make predictions, especially when these differ from the normatively acceptable ones.

The difference between statistics and psychology notwithstanding, during the course of this project we were several times surprised to see how close the two can be. Starting from a psychologically-motivated research on similarity, we axiomatized formulae that turned out to be identical to those used in kernel classification and kernel estimation of probabilities. We

likewise found ourselves axiomatizing the preference for theories with higher likelihood without ever suspecting that this is where the axioms would lead us. Thus, several times we asked ourselves which conditions it is likely to assume that real people satisfy, and we found that these conditions characterized well-known statistical techniques. As far as statistics is concerned, this means that some of these techniques, which were devised by statisticians without explicitly thinking in terms of axioms, ended up satisfying reasonable conditions. From a psychological point of view, these coincidences suggested that the human mind is probably a rather successful inference engine, in that general principles that make sense for human reasoning are also corroborated by statistical analysis.

These coincidences should not be overstated. There are surely many circumstances in which people tend to make silly predictions (cf. the Gambler's fallacy). On the other hand, there are many statistical techniques that are far from anything that can be viewed as a model of natural human reasoning. Moreover, the coincidences we find certainly do not imply that statistics and psychology are the same discipline. In fact, the opposite is true: statistics is mainly interested in developing techniques that are not obvious, namely, that go beyond the intuitive. By contrast, as far as inductive inference is concerned, psychology is interested in reasoning processes that tell us something new about the human mind, and these tend to correlate with less reasonable inferences.

It is therefore possible that the majority of real-life inferences are made by people in a very reasonable way that also corresponds to simple statistical techniques. Psychology would tend to focus on the remaining predictions that are not necessarily rational, and perhaps also not yet well understood. Statistics would ask how these predictions should be made, in ways that are too difficult for most people to come up with on their own. Thus, a sampling of real-life problems may suggest that the normative (statistics) is close to the descriptive (psychology), but a sampling of recent research in

either discipline would suggest the opposite. As the axiomatic approach we propose is rather rudimentary, it probably covers only the basic problems, where psychology and statistics may not be far apart. It is our hope that, while this domain covers a small portion of recent research, it does correspond to a non-negligible portion of everyday predictions.

4 The Combination Principle

4.1 Basic logic and results

Applying the axiomatic approach to the problem of inductive inference, we wish to identify reasonable patterns of inferences drawn from databases of observations. Thus, we do not focus on a single database and delve into the particular inferences that it entails. Rather, we consider an abstract method of inference, or a function that assigns sets of conclusions to databases, and ask which conditions should one impose on such a method or function.

In several studies we used axioms that are different manifestations of *the combination principle*: if a certain conclusion is reached given two disjoint databases, the same conclusion should be reached given their union. The precise meaning of "conclusion" and "database" should be specified. In fact, they are modeled in several ways in the papers presented here. A *database* can be a set of cases; or an ordered list of cases; or a counter vector, specifying how many times each type of case has been observed. In the first type of models, "union" is simply the union of two sets; in the second, "union" refers to concatenation of ordered lists of cases. Finally, if a database is no more than a counter vector, the union of two such databases corresponds to the addition of the two vectors, generating a new vector.

In general the three formulations differ and they may give rise to different operations on databases. However, in each of the axiomatic works that follow, we assumed that cases were exchangeable in an appropriate sense. If a database is a set, we define a notion of case-equivalence, and assume that

each particular case has infinitely many "replicas", that is, infinitely many cases that are equivalent to it. If a database is an ordered list of cases, the set of all cases already includes replicas because it includes lists in which the same case is repeated as many times as one wishes. In this context, we also assume that any list induces the same predictions as any permutation thereof. Finally, if a database is a vector of non-negative integers, counting how many times a case of each type has been encountered, specific cases do not explicitly appear, and exchangeability of cases is built into the model, as well as the assumption that each case can have as many replicas as we would like to consider.

Thus, in all three formulations we basically have in mind the same structure: there exist *types of cases*, and of each type we have observed a number of occurrence that is a non-negative integer. The order in which these cases were observed is immaterial. Should one wish, for example, that more recent cases would matter more than less recent ones, one could include the time of occurrence as one of the features of the case. The formal model, however, needs to assume that only the numbers of observations of each type matter.

It is a crucial and non-trivial assumption that cases can have as many replicas as one may imagine. For example, if one case is the financial crisis of 1929 and another – the crisis of 2008, one need not worry about the order in which they are listed, as each case contains enough information to describe its recency, and presumably its relevance. However, one may question how meaningful is it to consider a database in which the crisis of 1929 never occurred and that of 2008 occurred, say, 14 times. It is important to highlight that our axiomatic derivations do rely on the predictions that the reasoner would generate given each of the possible databases. In the case of global events such as wars, financial crises, and the like, the very formulation of the set of possible databases may lead us to question the relevance of the axiomatic derivation. By contrast, if one has in mind an application that is closer to cross sectional data, where different observations

are causally independent, it is not too demanding to assume that each case may appear any number of times, and to require that the reasoner should make a prediction given any number of occurrences of each case.

Next, we have to clarify what is meant by a *prediction*. In some papers, predictions are weak orders, ranking certain alternatives as at-least-as-likely-as others. The alternatives may be the possible values of the next observation, giving rise to models of frequency-based prediction, kernel classification, and kernel estimation. Alternatively, the alternatives may be general theories, or statistical models, which are ranked for plausibility given the data. This interpretation allows us to derive maximum likelihood-based selection of theories, as well as refinements thereof such as Akaike's information criterion, minimum-description-length criterion, etc. In other papers, the prediction is that a probability vector (or measure) lies on a certain line segment, or, equivalently, that a certain random variable has a pre-specified expectation.

The combination principle thus takes different shapes, depending on the model to which it applies. Its specific incarnations are referred to as "the combination axiom" or "the concatenation axiom", depending on the context. The axiomatic derivations make use of other axioms as well. These typically include an Archimedean condition and a richness condition. However, the conceptually important assumptions seem to be (i) that prediction is meaningfully defined for all databases; and that (ii) the combination principle holds.

Under these assumptions, our results are that there exists a function s over pairs of cases, which we tend to interpret as a similarity function, such that prediction can be represented by (the maximization of) s-weighted summation or by s-weighted averaging. More explicitly, assume that I is a counter vector, so that $I(c) \in \mathbb{Z}_+$ is the number of times cases (observations) of type c have been encountered. The set of case-types may be infinite, but it is assumed that $\sum_c I(c) < \infty$ for all databases I. Assume that, given I, $\succsim_I$ is a weak order on a set of alternatives. Gilboa and Schmeidler (2003a)

employs the combination principle, coupled with the other axioms, to derive the following similarity-weighted sum representation of $\succsim_I$: for each database I and any two alternatives a, b,

$$a \succsim_I b \qquad \Leftrightarrow \qquad \sum_c I(c)s(a,c) \geq \sum_c I(c)s(b,c). \tag{1}$$

Gilboa and Schmeidler (2010) modifies the combination principle to allow for a-priori biases for certain alternatives. Its leading interpretation is that elements such as a, b are theories, and these may differ in terms of their complexity. A preference for simpler theories may lead to a representation of the type

$$a \succsim_I b \qquad \Leftrightarrow \qquad w(a) + \sum_c I(c)s(a,c) \geq w(b) + \sum_c I(c)s(b,c) \tag{2}$$

which is axiomatized in this paper.

If predictions are more quantitative and take the form of probability vectors, Billot, Gilboa, Samet, and Schmeidler (2005) show that, as long as the domain of the prediction function is not limited to a single segment, the combination principle is equivalent to the following similarity-weighted averaging: for each case-type c there exists a number $s_c > 0$ and a probability vector p^c such that, for each I, the probability the reasoner chooses is

$$p(I) = \frac{\sum_c I(c)s_c p^c}{\sum_c I(c)s_c}.$$

An important special case of this rule is the following: one of finitely many states Ω will occur. Each past case c describes certain circumstances x_c and a realization of a state $\omega_c \in \Omega$. Given a new problem x_p, the probability of state ω is the similarity-weighted empirical frequency of ω in the past:

$$p(\omega|I) = \frac{\sum_c I(c)s(x_c, x_p)\mathbf{1}_{\{\omega_c=\omega\}}}{\sum_c I(c)s(x_c, x_p)}. \tag{3}$$

Clearly, this result does not apply if $|\Omega| = 2$, because in this case the range of the function I is included in a line segment. Gilboa, Lieberman, and Schmeidler (2006) provides an axiomatization of this formula in the two-outcome case.

4.2 Examples

There are several well-known statistical techniques that are special cases of the general representations above. Consider the simple similarity-weighted sum in (1), and suppose that the possible predictions (a, b) and past cases (c) belong to the same set, as in the case of a repeated roll of a die. Assume that the similarity function is the indicator function,

$$s(a, c) = \mathbf{1}_{\{a=c\}}$$

This means that only past occurrences of the very same prediction a may lend non-zero support to this prediction, and that all past cases are deemed equally relevant. In this case, the ranking according to (1) coincides with the ranking of possible predictions by their frequency in the past.

Next assume that the set to which past cases c and possible predictions a belong is infinite, such as $\mathbb{R}^k$. In this case it makes sense to allow the function $s(a, c)$ to be positive also when a and c are close, though not necessarily identical. Then, the expressions on the right hand side of (1) are those used for kernel estimation of a density function (see Akaike, 1954, Rosenblatt, 1956, Parzen, 1962, Silverman, 1986, and Scott, 1992, for a survey).

Along similar lines, assume that the prediction problem is a classification problem: each observation $c = (x_c, a_c)$ consists of data x_c (say, a point in $\mathbb{R}^k$) and a class a_c to which the point is known to belong. Given a new point with parameters x, the reasoner is asked to guess to which class a it belongs. Then, specifying

$$s\left(a, (x_c, a_c)\right) = k\left(x_c, x\right) \mathbf{1}_{\{a=a_c\}}$$

the formula (1) boils down to kernel classification with a kernel function k.

More interesting examples involve applications where the set of observations and the possible predictions have no common structure. For example, if cases c are past observations, and the predictions a are theories, they do not typically belong to the same set. However, the axiomatization suggests that, if the reasoner satisfies the axioms, one can find a function s that would

describe the reasoner's predictions via (1) and thus, indirectly, reflect the reasoner's perception of the relationship between theories and observations. Specifically, assume that the function s is negative.[2] Define, for a theory a and an observation c,

$$p(c|a) = \exp(s(a,c))$$

so that

$$\log(p(c|a)) = s(a,c)$$

and (1) becomes equivalent to ranking of theories by the (log-)likelihood function.

The introduction of a-priori biases to the ranking of theories, such as the preference for simplicity, suggests that a constant should be added to the log-likelihood function as in (2). Clearly, this formulation includes as special cases the Akaike Information Criterion (AIC, Akaike, 1974) and Minimum Description Length criterion for model selection (MDL, see Wallace and Boulton, 1968, Wallace and Dowe, 1999, and Wallace, 2005 for a more recent survey). These do not fall under the category of (1), where there is no room for the function w. Indeed, ranking of theories by AIC or by MDL does not satisfy the combination principle as stated. Gilboa and Schmeidler (2010) invoke a weaker version of this principle to derive such ranking rules.

4.3 Limitations

The combination principle appears to be rather intuitive, and it is perhaps not too surprising that this principle is satisfied by a variety of statistical techniques. Yet, there are also many statistical techniques, as well as natural reasoning procedures, that do not satisfy it. These violations can be classified into three types: first, there are situations in which the principle is inappropriately applied. Given the generality of the principle, stated for general "conclusions" drawn from databases, there are situations where the principle

[2] The analysis in Gilboa and Schmeidler (2003) shows that this assumption involves no loss of generality as long as the number of theories is finite.

may formally apply, yet it may not be very sensible to adhere to it. For example, Simpson's paradox (see Simpson, 1951, or, for example, de Groot, 1975) is a well-known example in which a certain conclusion can be drawn from each of two databases but not from their union. As we argue in Gilboa and Schmeidler (2010), this is a mis-application of the principle, because in this example the theories concerned are not directly about the single data, but about certain patterns in the data. More specifically, the completeness axiom (which, in one shape or another, appears in all axiomatizations mentioned here) expects one to rank theories given each and every database, even if the database contains only one observation. This does not seem to be the case in Simpson's paradox, and thus we argue that this violation of the combination principle is due to a mis-application of the model. Put differently, the completeness axiom implicitly restricts the type of observation-prediction pairs to which the theory should be applied.

A second class of violations of the combination principle are those in which one considers a statistical technique and concludes that it is indeed a theoretical flaw that it fails to satisfy such a reasonable principle. According to our personal taste, this is the case with k-nearest neighbor techniques (Fix and Hodges, 1951, 1952, Cover and Hart, 1967), which violate the combination axiom (as in Gilboa Schmeidler, 2003a) because the weight assigned to an observation does not depend solely on its inherent relevance, but also on its relative relevance, as compared to other observations. While this is ultimately a subjective judgment, we find that this violation is not among the merits of k-nearest neighbor techniques.

Finally, there are violations of the third type, in which one finds that the principle is too restrictive. Two main such categories are situations in which one learns the similarity function from the data, and when one engages in combination of induction and deduction. We view each of these as pointing to important directions for future research, and discuss them separately.

5 Learning the Similarity

Case-based reasoning relies on the similarity that one finds between cases. Where does this similarity function come from? Taking a descriptive interpretation, this question brings us to the domain of psychology, and it would suggest that the similarity function is not a fixed, immutable reasoning tool that one is born with. Rather, it is learnt from experience. For example, a physician may learn, through her experience, that for a particular diagnosis, weight and blood pressure are important features of similarity, whereas blood type is not. Thus, past cases do not only suggest what will be the outcome of future cases using the similarity function; they also indicate which similarity functions are better suited to perform this type of case-to-case induction.

Taking a normative viewpoint, closer to the statistical mindset, it stands to reason that one may update the similarity function based on data. Indeed, kernel methods typically change the kernel function as data accumulate, so that in larger databases a tighter kernel is used, allowing the prediction to be based mostly on the more relevant cases when there are sufficiently many of these. But beyond the sheer number of past cases, their content can also serve as a guide regarding the choice of the similarity/kernel function.

In Gilboa, Lieberman, and Schmeidler (2006) we formally introduce the notion of *empirical similarity*. This is defined as a similarity function that, within a pre-specified class of functions, minimizes the sum of squared errors one would have obtained, were one to use that similarity function in the past, predicting each outcome based on the rest of the database (the "leave-one-out" criterion). We develop the statistical theory for estimation of the similarity function, assuming that the process is indeed governed by a similarity-weighted-average of other (or past) observations.

The empirical similarity idea can also be used to analyze databases that are not necessarily believed to have been generated by a similarity-based process. Rather, one can suggest the idea as a statistical prediction technique that mimics the informal learning of similarity that human beings

naturally engage in. Further, one can follow this line of reasoning and use the empirical similarity to define objective probabilities: probabilities are defined by similarity-weighted frequencies in past cases, where the similarity function is learnt from the same database. Thus, one can shed the subjective baggage of the psychological notion of similarity, and replace it by a notion of empirical similarity, which has a claim to objectivity similar to those of other statistical constructs. Gilboa, Lieberman, and Schmeidler (2009) is devoted to this definition of objective probabilities, and it also discusses the tension between the proposal to learn the similarity function and the combination principle that is violated by such learning.

6 Rule-Based Reasoning

We originally formulated the combination principle with case-to-case induction in mind. Somewhat to our surprise, we found that it can be reinterpreted for case-to-rule (or observation-to-theory) induction, and that it then basically coincides with maximum likelihood selection of models. Further, as mentioned above, the principle can be adapted to introduce considerations other than the likelihood function, such as simplicity or prior probability, into model selection.

However, when theories are selected based on past observations, and they are then used to forecast future observations, the combination principle does not seem appropriate. For example, if one uses observations of variables x, y in order to estimate a regression model $y = \alpha + \beta x + \varepsilon$, the selection of a model (or "theory" or "rule") boils down to the selection of the parameters α, β. Under the standard assumptions, least square estimators are maximum likelihood estimators, and the ranking of parameter values by the likelihood function will satisfy the combination principle. However, if the selected parameters are then used, via the regression equation, to predict the value of y for a new observation x, the combination principle will be violated.

Moreover, this should be expected to be the case whenever one engages in combined inductive and deductive reasoning: inductive reasoning to find a model based on the data, and deductive reasoning to predict the data based on the selected model.

Thus, we find that the theory developed here has little to say about rule-based prediction. Moreover, Aragones, Gilboa, Postlewaite, and Schmeidler (2005) shows that the theory selection problem is a computational hard one. Specifically, when one introduces goodness of fit as well as simplicity as model selection criteria, very reasonable formulations of the problem render the selection of the "best" theory an NP-Hard problem. This implies that even the first step, of case-to-rule induction, may be too complicated to be performed, either by humans or by computers.

7 Summary

We believe that the axiomatic approach to inductive inference is important and useful. It helps us understand what theories actually assume; it highlights equivalences between different formulations of the same theory; it guarantees that theories have a clear empirical content; and it may ensure that the method one chooses for inductive reasoning is coherent and sensible even if the database is small and asymptotic results are of limited relevance. Ideally, one would like to have axiomatic derivations of all theories one uses, and use the axioms to help select a theory for specific classes of inductive inference problems.

Unfortunately, the results we report here indicate only partial success. All the axiomatic results rely very heavily on the combination principle. We find this principle a reasonable starting point, but it certainly cannot be considered a universal condition on inductive inference. Future research might find more flexible axiomatic approaches that would be able to generalize the theories presented here to include other types of reasoning.

The book is organized as follows. As background, we start with two axiomatic derivations of case-based decision theory. The first, Gilboa and Schmeidler (1995), is the original paper, highlighting the basic ideas. The second, Gilboa and Schmeidler (1997), extends the theory to incorporate act-similarity considerations, and introduces the mathematical structure of the combination principle. Several of the subsequent papers use this paper as their mathematical backbone.

Next we consider inductive inference as modeled by an "at least as likely as" relation. The basic tools are given in Gilboa and Schmeidler (2003a). If the objects to be ranked are events in a given state space, one may hope to say more, as there is a measure-theoretic structure one may use. On the other hand, some auxiliary assumptions are inappropriate in this context. Gilboa and Schmeidler (2002) deals with this case, and with the combinatorial issues that arise if one wishes to obtain a probability function over events based on likelihood rankings that satisfy the combination principle.

The application of Gilboa and Schmeidler (2003a) to theory selection is limited to the maximum likelihood principle. As mentioned above, this is extended to an additive trade-off between simplicity and likelihood in Gilboa and Schmeidler (2010). This paper is therefore the next in the volume.

When the observable data are numerical probabilities, Billot, Gilboa, Samet, and Schmeidler (2005) provide an axiomatization of the similarity-weighted-frequency formula. It is followed by Billot, Gilboa, and Schmeidler (2008), which characterizes the exponential similarity function in the context of this formula.

However, one may only go so far when using theoretical considerations for the selection of a similarity function. In the final analysis, the choice of the function remains an empirical issue, which is what Gilboa, Lieberman, and Schmeidler (2006) is about. This paper also completes the axiomatization of the similarity-weighted frequency formula for the single-dimensional case. It is followed by Gilboa, Lieberman, and Schmeidler (2009), which offers

the empirical similarity, coupled with the similarity-weighted formula, as a definition of objective probabilities.

While the empirical similarity papers suggest a new statistical technique and a new definition of objective probabilities, they also highlight the limitation of the axiomatizations provided here, as they focus on violating the combination principle (which lies at the heart of these axiomatizations). We then move to discuss the complexity of theory selection, in Aragones, Gilboa, Postlewaite, and Schmeidler (2005), which indicates another important direction in which the axiomatic theory of inductive inference may be enriched.

Finally, we conclude with two applications of the mathematical techniques developed in Gilboa and Schmeidler (1997, 2003) to other problems. Gilboa and Schmeidler (2003b) applies the method to the derivation of a utility function, coupled with the expected utility principle, in the context of a game, that is, without referring to lotteries other than the game offers. Gilboa and Schmeidler (2004) offers a definition of subjective probabilities limited to the distributions of given random variables, without reference to the (much larger) underlying state space. None of these two papers is related to the main project in terms of content. However, both contain results that may be useful for certain extensions, such as modelling memory in a continuous way.

References

Akaike, H. (1954), "An Approximation to the Density Function", *Annals of the Institute of Statistical Mathematics*, **6**: 127-132.

——— (1974), "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* **19** (6), 716–723.

Alchourron, C.E., P. Gardenfors, and D. Makinson (1985), "On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision," *Journal of Symbolic Logic,* **50**: 510–530.

Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler (2005), "Fact-Free Learning", *American Economic Review*, **95**: 1355-1368.

Bewley, T. (2002), "Knightian Decision Theory: Part I", *Decisions in Economics and Finance*, 25: 79-110.

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2005), "Probabilities as Similarity-Weighted Frequencies", *Econometrica*, **73**: 1125-1136.

Billot, A., I. Gilboa, and D. Schmeidler (2008), "Axiomatization of an Exponential Similarity Function", *Mathematical Social Sciences*, **55**: 107–115.

Carnap, R. (1923), "Uber die Aufgabe der Physik und die Andwednung des Grundsatze der Einfachstheit", *Kant-Studien*, **28**: 90-107.

Choquet, G. (1953), "Theory of Capacities", *Annales de l'Institut Fourier,* 5: 131-295.

Cover, T. and P. Hart (1967), "Nearest Neighbor Pattern Classification", IEEE *Transactions on Information Theory* **13**: 21-27.

de Finetti, B. (1931), Sul Significato Soggettivo della Probabilità, *Fundamenta Mathematicae*, 17: 298-329.

——— (1937), "La Prevision: ses Lois Logiques, ses Sources Subjectives", *Annales de l'Institut Henri Poincare*, 7: 1-68 (translated in *Studies in Subjective Probability*, edited by H.E. Kyburg and H.E. Smokler, Wiley, 1963).

de Groot, M. H. (1975), *Probability and Statistics*, Reading, MA: Addison-Wesley Publishing Co.

Ellsberg, D. (1961), "Risk, Ambiguity and the Savage Axioms", *Quarterly Journal of Economics*, 75: 643-669.

Fix, E. and J. Hodges (1951), "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties". Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

——— (1952), "Discriminatory Analysis: Small Sample Performance". Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Gardenfors, P. (1992), *Belief Revision: An Introduction.* Cambridge: Cambridge University press.

Gilboa, I. and M. Marinacci (2010), "Ambiguity and the Bayesian Paradigm", mimeo.

Gilboa, I. and D. Schmeidler (1989), "Maxmin Expected Utility with a Non-Unique Prior", *Journal of Mathematical Economics*, 18: 141-153.

——— (1995), "Case-Based Decision Theory", *Quarterly Journal of Economics*, **110**: 605-639.

——— (1997), "Act Similarity in Case-Based Decision Theory", *Economic Theory*, **9**: 47-61.

——— (2001), *A Theory of Case-Based Decisions.* Cambrdige: Cambridge University Press.

——— (2002), "A Cognitive Foundation of Probability", *Mathematics of Operations Research*, **27**: 68-81.

——— (2003a), "Inductive Inference: An Axiomatic Approach", *Econometrica*, **71**: 1-26.

——— (2003b), "Expected Utility in the Context of a Game", *Games and Economic Behavior*, **44**: 184-194.

——— (2004), "Subjective Distributions", *Theory and Decision*, **56**: 345-357.

——— (2010), "Likelihood and Simplicity: An Axiomatic Approach", *Journal of Economic Theory*, **145**: 1757-1775.

Gilboa, I., O. Lieberman, and D. Schmeidler (2006), "Empirical Similarity", *Review of Economics and Statistics*, **88**: 433-444.

——— (2009), "On the Definition of Objective Probabilities by Empirical Similarity", *Synthese*.

Hume, D. (1748), *Enquiry into the Human Understanding*. Oxford, Clarendon Press.

Keynes, J. M. (1921), *A Treatise on Probability*. London: MacMillan and Co.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005), "A Smooth Model of Decision Making under Ambiguity," *Econometrica*, 73: 1849-1892.

Knight, F. H. (1921), *Risk, Uncertainty, and Profit*. Boston, New York: Houghton Mifflin.

Maccheroni, F., M. Marinacci, and A. Rustichini (2006), "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences," *Econometrica*, 74: 1447-1498.

Nau, R.F. (2006), "Uncertainty Aversion with Second-Order Utilities and Probabilities", *Management Science* 52: 136-145.

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Parzen, E. (1962), "On the Estimation of a Probability Density Function and the Mode", *Annals of Mathematical Statistics*, **33**: 1065-1076.

Popper, K.R. (1934), *Logik der Forschung;* English edition (1958), *The Logic of Scientific Discovery.* London: Hutchinson and Co. Repreinted (1961), New York: Science Editions.

Ramsey, F. P. (1931), "Truth and Probability", *The Foundation of Mathematics and Other Logical Essays.* New York: Harcourt, Brace and Co.

Riesbeck, C. K. and R. C. Schank (1989), *Inside Case-Based Reasoning.* Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function", *Annals of Mathematical Statistics*, **27**: 832-837.

Royall, R. (1966), *A Class of Nonparametric Estimators of a Smooth Regression Function.* Ph.D. Thesis, Stanford University, Stanford, CA.

Savage, L. J. (1954), *The Foundations of Statistics.* New York: John Wiley and Sons.

Schank, R. C. (1986), *Explanation Patterns: Understanding Mechanically and Creatively.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmeidler, D. (1986), "Integral Representation without Additivity." *Proceedings of the American Mathematical Society*, 97: 255-261.

——— (1989), "Subjective Probability and Expected Utility without Additivity", *Econometrica*, 57: 571-587.

Seo, K. (2009), "Ambiguity and Second-Order Belief", *Econometrica*, **77**: 1575-1605.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: John Wiley and Sons.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis.* London and New York: Chapman and Hall.

Simpson, E. H. (1951). "The Interpretation of Interaction in Contingency Tables". *Journal of the Royal Statistical Society, Ser. B* **13**: 238–241.

Tversky, A. and D. Kahneman (1974), "Judgment under Uncertainty: Heuristics and Biases", *Science*, **185**: 1124-1131.

Wallace, C.S. (2005), *Statistical and Inductive Inference by Minimum Message Length* Series: Information Science and Statistics, Springer.

Wallace, C.S. and D. M. Boulton (1968), "An Information Measure for Classification", *Comput. J.*, **13**, 185-194.

Wallace, C. S. and D. L. Dowe (1999), "Minimum Message Length and Kolmogorov Complexity", *The Computer Journal*, **42**, 270-283.