Inductive Inference: An Axiomatic Approach*

Itzhak Gilboa† and David Schmeidler‡

July 1999 – Revised February 2001

Abstract

A predictor is asked to rank eventualities according to their plausibility, based on past cases. We assume that she can form a ranking given any memory that consists of finitely many past cases. Mild consistency requirements on these rankings imply that they have a numerical representation via a matrix assigning numbers to eventuality-case pairs, as follows. Given a memory, each eventuality is ranked according to the sum of the numbers in its row, over cases in memory. The number attached to an eventuality-case pair can be interpreted as the degree of support that the past lends to the plausibility of the eventuality. Special cases of this result may be viewed as axiomatizing kernel methods for estimation of densities and for classification problems. Interpreting the same result for rankings of theories or hypotheses, rather than of specific eventualities, it is shown that one may ascribe to the predictor subjective conditional probabilities of cases given theories, such that her rankings of theories agree with rankings by the likelihood functions.

*We wish to thank Yoav Binyamini, Didier Dubois, Drew Fudenberg, Bruno Jullien, Edi Karni, Simon Kasif, Daniel Lehmann, Sujoy Mukerji, Roger Myerson, Klaus Nehring, Ariel Rubinstein, Lidror Troyanski, Peter Wakker, Peyton Young, and two anonymous referees for the discussions that motivated this work, as well as for comments and references. Most of this material is also available as a Web paper at http://econ.tau.ac.il/gilboa/Inductive_Inference/Index.html.

†Tel-Aviv University. igilboa@post.tau.ac.il

‡Tel-Aviv University and The Ohio State University. schmeid@post.tau.ac.il

1 Introduction

Prediction is based on past cases. As Hume (1748) argued, "From causes which appear similar we expect similar effects. This is the sum of all our experimental conclusions." Over the past decades Hume's approach has found re-incarnations in the artificial intelligence literature as reasoning by analogies, reasoning by similarities, or case-based reasoning. (See Schank (1986) and Riesbeck and Schank (1989).) Many authors accept the view that analogies, or similarities to past cases hold the key to human reasoning. Moreover, the literature on machine learning and pattern recognition deals with using past cases, or observations, for predicting or classifying new data. (See, for instance, Forsyth and Rada (1986) and Devroye, Gyorfi, and Lugosi (1996).) But how should past cases be used? How does, and how should one resolve conflicts between different analogies? To address these questions, let us first consider a few examples.

Example 1: A die is rolled over and over again. One has to predict the outcome of the next roll. As far as the predictor can tell, all rolls were made under identical conditions. Also, the predictor does not know of any a-priori reason to consider any outcome more likely than any other. The most reasonable prediction seems to be the mode of the empirical distribution, namely, the outcome that has appeared most often in the past. Moreover, empirical frequencies suggest a plausibility ranking of all possible outcomes, and not just a choice of the most plausible ones.[1]

Example 2: A physician is asked by a patient if she predicts that a surgery will succeed in his case. The physician knows whether the procedure succeeded in most cases in the past, but she will be quick to remind her patient that every human body is unique. Indeed, the physician knows that

[1]The term "likelihood" in the context of a binary relation, "at least as likely as", has been used by de Finetti (1937) and by Savage (1954). It should not be confused with "likelihood" in the the context of likelihood functions, also used in the sequel. At this point we use "likelihood" and "plausibility" informally and interchangeably.

the statistics she read included patients who varied in terms of age, gender, medical condition, and so forth. It would therefore be too naive of her to quote statistics as if the empirical frequencies were all that mattered. On the other hand, if the physician considers only past cases of patients that are identical to hers, she will probably end up with an empty database.

Example 3: An expert on international relations is asked to predict the outcome of the conflict in Kosovo. She is expected to draw on her vast knowledge of past cases, coupled with her astute analysis thereof, in forming her prediction. As in Example 2, the expert has a lot of information she can use, but she cannot quote even a single case that was identical to the situation at hand. Moreover, as opposed to Example 2, even the possible eventualities are not identical to outcomes that occurred in past cases.

We seek a theory of prediction that will make use of the available information, but will allow different past cases to have differential relevance to the prediction problem. Specifically, we consider a prediction problem for which a set of possible eventualities is given. This set may or may not be an exhaustive list of all conceivable eventualities. We do not model the process by which such a set is generated. Rather, we assume the set given and restrict attention to the problem of qualitative ranking of its elements according to their likelihood.

The prediction rule Consider the following prediction rule, say, for Example 2. The physician considers all known cases of successful surgery. She uses her subjective judgment to evaluate the similarity of each of these cases to the patient she is treating, and she adds them up. She then does the same for unsuccessful treatments. It seems reasonable that the outcome with the larger aggregate similarity value will be her prediction. This generalizes frequentist ranking to a "fuzzy sample": in both examples, likelihood of an outcome is measured by summation over cases in which it occurred. Whereas in Example 1 the weight attached to each past case is 1, in this example this weight varies according to the physician's subjective assessment of similarity

of the relevant cases. Rather than a dichotomous distinction between data points that do and those that do not belong to the sample, each data point belongs to the sample to a certain degree, say, between 0 and 1.

The prediction rule we propose can also be applied to Example 3 as follows. For each possible outcome of the conflict in Kosovo, and for each past case, the expert is asked to assess a number, measuring the degree of support that the case lends to this outcome. Adding up these numbers, for all known cases and for each outcome, yields a numerical representation of the likelihood ranking. Thus, our prediction rule can be applied also when there is no structural relationship between past cases and future eventualities.

Formally, let M denote the set of known cases. For each $c \in M$ and each eventuality x, let $v(x, c) \in \mathbb{R}$ denote the degree of support that case c lends to eventuality x. Then the prediction rule ranks eventuality x as more likely than eventuality y if and only if

$$(\circ) \qquad \sum_{c \in M} v(x, c) > \sum_{c \in M} v(y, c).$$

Axiomatization The main goal of this paper is to axiomatize this rule. We assume that a predictor has a ranking of possible eventualities given any possible memory (or database). A memory consists of a finite set of past cases, or stories. The predictor need not envision all possible memories. She might have a rule, or an algorithm that generates a ranking (in finite time) for each possible memory. We only rely on qualitative plausibility rankings, and do not assume that the predictor can quantify them in a meaningful way. Cases are not assumed to have any particular structure. However, we do assume that for every case there are arbitrarily many other cases that are deemed equivalent to it by the predictor (for the prediction problem at hand). For instance, if the physician in Example 2 focuses on five parameters of the patient in making her prediction, we can imagine that she has seen arbitrarily many patients with particular values of the five parameters. The equivalence relation on cases induces an equivalence relation on memories (of equal sizes), and the latter allows us to consider replication of memories,

that is, the disjoint union of several pairwise equivalent memories.

Our main assumption is that prediction satisfies a *combination axiom*. Roughly, it states that if an eventuality x is more likely than an eventuality y given two possible disjoint memories, then x is more likely than y also given their union. For example, assume that the patient in Example 2 consults two physicians, who were trained in the same medical school but who have been working in different hospitals since graduation. Thus, the physicians can be thought of as having disjoint databases on which they can base their prediction, while sharing the inductive algorithm. Assume next that both physicians find that success is more likely than failure in the case at hand. Should the patient ask them to share their databases and re-consider their predictions? If the inductive algorithm that the physicians use satisfies the combination axiom, the answer is negative.

We also assume that the predictor's ranking is *Archimedean* in the following sense: if a database M renders eventuality x more likely than eventuality y, then for every other database N there is a sufficiently large number of replications of M, such that, when these memories are added to N, they will make eventuality x more likely than eventuality y. Finally, we need an assumption of *diversity*, stating that any list of four eventualities may be ranked, for some conceivable database, from top to bottom. Together, these assumptions necessitate that prediction be made according to the rule suggested by the formula $(\circ)$ above. Moreover, we show that the function v in $(\circ)$ is essentially unique.

This result can be interpreted in several ways. From a descriptive viewpoint, one may argue that experts' predictions tend to be consistent as required by our axioms (of which the combination is the most important), and that they can therefore be represented as aggregate similarity-based predictions. From a normative viewpoint, our result can be interpreted as suggesting the aggregate similarity-based predictions as the only way to satisfy our consistency axioms. In both approaches, one may attempt to measure

similarities using the likelihood rankings given various databases.

Observe that we assume no a priori conceptual relationship between cases and eventualities. Such relationships, which may exist in the predictor's mind, will be revealed by her plausibility rankings. Further, even if cases and eventualities are formally related (as in Example 2), we do not assume that a numerical measure of distance, or of similarity is given in the data.

Axiomatization of kernel methods A well-known statistical problem is the estimation of a density function, based on a finite sample of observations. In this case, a common statistical technique is kernel estimation (see Akaike (1954), Rosenblatt (1956), Parzen (1962), and Silverman (1986) and Scott (1992) for recent texts): a kernel function $k(x, y)$ is chosen, and the estimated density, based on observations $\{x_i\}_{i=1}^n$, is $f(y) = \sum_{i=1}^n k(x_i, y)$. In other words, every observation x_i is assumed to make every y in its vicinity (as defined by the kernel function k) more likely.

Kernel estimation of density functions is a special case of our model. Thus, when our model is applied to the special structure above, it can be viewed as axiomatizing this estimation method. This may serve as a normative justification for this method, as well as a definition of the kernel function in terms of qualitative plausibility rankings.

Kernel methods have also been applied to classification problems. In such a problem a classifier is equipped with a set of data points for which the correct class is given, and it is asked to classify the next data point. Kernel methods employ a kernel function defined on pairs of data points, and, for each possible class, compute the sum, over all examples of this class, of the kernel function between these known examples and the new data point. A maximizer of this sum is chosen as the classification of the new data point.

Clearly, kernel classification methods are a special case of our prediction rule. Indeed, our description of the prediction rule in Example 2 can be viewed as a generalization of kernel classification to a situation where the data need not be points in a Euclidean space, and where distance, similarity,

or kernel functions are not given a-priori.

When applied to the special case of classification problems, our result can be used to derive an axiomatization of kernel methods. This can be used to justify these methods, and to define a subjective kernel function based on qualitative plausibility rankings.

Axiomatization of maximum likelihood Prediction is not restricted to single cases. Often one is asked to choose not only the most plausible outcome in a given instance, but the most plausible theory, or hypothesis. How should we use cases in this problem?

We argue that our axioms are reasonable for this case as well. Indeed, the main axiom is that, if, based on each of two disjoint databases, we tend to prefer theory T to theory T', we should have the same preference based on the union of the databases. Hence our rule is also a reasonable suggestion for ranking theories given data: every theory and every case are ascribed a number, and the plausibility of a theory is measured by the summation of the numbers corresponding to it, over all cases in memory.

Let us suppose that the numbers assigned to theory-case pairs are negative. They can then be viewed as the logarithms of the conditional probabilities of cases given theories. Summing up these numbers over all cases corresponds to multiplying the conditional probabilities. In other words, the numerical function that measures the plausibility of a theory is simply the log-likelihood function.

Thus our theorem can be viewed as an axiomatization of likelihood ranking. Given a qualitative "at least as plausible as" relation between theories we derive numerical conditional probabilities for each theory and each case, together with the algorithm that ranks theories based on their likelihood function, under the assumption that all cases were statistically independent. While the conditional probabilities we derive are unique only up to certain transformations, our result does provide a compelling reason to use likelihood rankings.

Methodological remarks The Bayesian approach (Ramsey (1931), de Finetti (1937), and Savage (1954)) holds that all prediction problems should be dealt with by a prior subjective probability that is updated in light of new information via Bayes' rule. This requires that the predictor have a prior probability over a space that is large enough to describe all conceivable new information. We find that in certain examples (as above) this assumption is not cognitively plausible. By contrast, the prediction rule ($\circ$) requires the evaluation of support weights only for cases that were actually encountered. For an extensive methodological discussion, see Gilboa and Schmeidler (2001).

Since the early days of probability theory, the concept of probability serves a dual role: one relating to empirical frequencies, and the other – to quantification of subjective beliefs or opinions. (See Hacking (1975).) The Bayesian approach offers a unification of these roles employing the concept of a subjective prior probability. Our approach may also be viewed as an attempt to unify the notions of empirical frequencies and subjective opinions. Whereas the axiomatic derivations of de Finetti (1937) and Savage (1954) treat the process of the generation of a prior as a black box, our rule aims to make a preliminary step towards the modeling of this process.

Thus, our approach is complementary to the Bayesian approach at two levels: first, it may offer an alternative model of prediction, when the information available to the predictor is not easily translated to the language of a prior probability. Second, our approach may describe how a prior is generated. (See also Gilboa and Schmeidler (1999).)

The rest of this paper is organized as follows. Section 2 presents the formal model and the main results. Section 3 discusses the relationship to kernel methods and to nearest neighbor approaches. Section 4 discusses the derivation of maximum likelihood rankings. Section 5 contains a critical discussion of the axioms, attempting to outline their scope of application. Finally, Section 6 briefly discusses alternative interpretations of the model, and, in

particular, relates it to case-based decision theory. Proofs are relegated to the appendix.

2 Model and Result

2.1 The framework

The primitives of our model consist of two non-empty sets X and $\mathbb{C}$. We interpret X as the set of all conceivable *eventualities* in a given prediction problem, p, whereas $\mathbb{C}$ represents the set of all conceivable *cases*. To simplify notation, we suppress the prediction problem p whenever possible. The predictor is equipped with a finite set of cases $M \subset \mathbb{C}$, her *memory*, and her task is to rank the eventualities by a binary relation, "at least as likely as".

While evaluating likelihoods, it is insightful not only to know what has happened, but also to take into account what could have happened. The predictor is therefore assumed to have a well-defined "at least as likely as" relation on X for many other collections of cases in addition to M itself. Let $\mathbb{M}$ be the set of finite subsets of $\mathbb{C}$. For every $M \in \mathbb{M}$, we denote the predictor's "at least as likely as" relation by $\succsim_M \subset X \times X$.

Two cases c and d are *equivalent*, denoted $c \sim d$, if, for every $M \in \mathbb{M}$ such that $c, d \notin M$, $\succsim_{M\cup\{c\}} = \succsim_{M\cup\{d\}}$. To justify the term, we note the following.

Observation: $\sim$ is an equivalence relation.

Note that equivalence of cases is a subjective notion: cases are equivalent if, in the eyes of the predictor, they affect likelihood rankings in the same way. Further, the notion of equivalence is also context-dependent: two cases c and d are equivalent as far as a specific prediction problem is concerned.

We extend the definition of equivalence to memories as follows. Two memories $M_1, M_2 \in \mathbb{M}$ are equivalent, denoted $M_1 \sim M_2$, if there is a bijection $f : M_1 \to M_2$ such that $c \sim f(c)$ for all $c \in M_1$. Observe that memory equivalence is also an equivalence relation. It also follows that, if $M_1 \sim M_2$, then, for every $N \in \mathbb{M}$ such that $N \cap (M_1 \cup M_2) = \varnothing$, $\succsim_{N\cup M_1} = \succsim_{N\cup M_2}$.

Throughout the discussion, we impose the following structural assumption.

Richness Assumption: For every case $c \in \mathbb{C}$, there are infinitely many cases $d \in \mathbb{C}$ such that $c \sim d$.

A note on nomenclature: the main result of this paper is interpreted as a representation of a prediction rule. Accordingly, we refer to a "predictor" who may be a person, an organization, or a machine. However, the result may and will be interpreted in other ways as well. Instead of ranking eventualities one may rank *decisions*, *acts,* or a more neutral term, *alternatives*. Cases, the elements of $\mathbb{C}$, may also be called *observations* or *facts*. A memory M in $\mathbb{M}$ represents the predictor's *knowledge* and will be referred to also as a *database*.

2.2 The axioms

We will use the four axioms stated below. In their formalization let $\succ_M$ and $\approx_M$ denote the asymmetric and symmetric parts of $\succsim_M$, as usual. $\succsim_M$ is *complete* if $x \succsim_M y$ or $y \succsim_M x$ for all $x, y \in X$.

A1 Order: For every $M \in \mathbb{M}$, $\succsim_M$ is complete and transitive on X.

A2 Combination: For every disjoint $M, N \in \mathbb{M}$ and every $x, y \in X$, if $x \succsim_M y$ ($x \succ_M y$) and $x \succsim_N y$, then $x \succsim_{M \cup N} y (x \succ_{M \cup N} y)$.

A3 Archimedean Axiom: For every disjoint $M, N \in \mathbb{M}$ and every $x, y \in X$, if $x \succ_M y$, then there exists $l \in \mathbb{N}$ such that for any l-list $(M_i)_{i=1}^{l}$ of pairwise disjoint M_i's in $\mathbb{M}$, where for all $i \leq l$, $M_i \sim M$ and $M_i \cap N = \emptyset$, $x \succ_{M_1 \cup \ldots \cup M_l \cup N} y$ holds.

Axiom 1 simply requires that, given any conceivable memory, the predictor's likelihood relation over eventualities is a weak order. Axiom 2 states that if eventuality x is more plausible than eventuality y given two disjoint memories, x should also be more plausible than y given the union of these memories. Axiom 3 is states that if, given the memory M, the predictor

believes that eventuality x is strictly more plausible than y, then, no matter what is her ranking for another memory, N, there is a number of "repetitions" of M that is large enough to overwhelm the ranking induced by N.

Finally, we need a diversity axiom. It is not necessary for representation of likelihood relations by summation of real numbers. Theorem 1 below is an equivalence theorem, characterizing precisely which matrices of real numbers will satisfy this axiom.

A4 Diversity: For every list (x, y, z, w) of distinct elements of X there exists $M \in \mathbb{M}$ such that $x \succ_M y \succ_M z \succ_M w$. If $|X| < 4$, then for any strict ordering of the elements of X there exists $M \in \mathbb{M}$ such that $\succ_M$ is that ordering.

2.3 The main results

For clarity of exposition, we first formulate the key result.

Result: Let there be given X, $\mathbb{C}$, and $\{\succsim_M\}_{M\in\mathbb{M}}$ satisfying the richness assumption and A1-A4. Then there is a matrix $v : X \times \mathbb{C} \to \mathbb{R}$ such that:

$$(*) \quad \begin{cases} \text{for every } M \in \mathbb{M} \text{ and every } x, y \in X, \\ \\ x \succsim_M y \quad \text{iff} \quad \sum_{c\in M} v(x,c) \geq \sum_{c\in M} v(y,c) \ . \end{cases}$$

In other words, axioms A1-A4 imply that $\{\succsim_M\}_{M\in\mathbb{M}}$ follow our prediction rule for an appropriate choice of the matrix v. Not all of these axioms are, however, necessary for the representation to obtain. Indeed, the axioms imply special properties of the representing matrix v. First, it can be chosen in such a way that equivalent cases are attached identical columns. Second, every four rows of the matrix satisfy an additional condition. Existence of a matrix v satisfying these two properties together with $(*)$ does imply axioms A1-A4. Finally, the matrix v is essentially unique. Theorem 1 below states the exact characterization and uniqueness results. Before stating the theorem, we present two additional definitions.

Definition: A matrix $v : X \times \mathbb{C} \to \mathbb{R}$ *respects case equivalence* (with respect to $\{\succsim_M\}_{M\in\mathbb{M}}$) if for every $c, d \in \mathbb{C}$, $c \sim d$ iff $v(\cdot, c) = v(\cdot, d)$.

When no confusion is likely to arise, we will suppress the relations $\{\succsim_M\}_{M\in\mathbb{M}}$ and will simply say that "v respects case equivalence".

The following definition applies to real-values matrices in general. It will be used for the matrix $v : X \times \mathbb{C} \to \mathbb{R}$ in the statement of the theorem, but also for another matrix in the proof. It defines a matrix to be diversified if no row in it is dominated by an affine combination of any other three (or less) rows. Thus, if v is diversified, no row in it dominates another. Indeed, the property of diversification can be viewed as a generalization of this condition.
Definition: A matrix $v : X \times Y \to \mathbb{R}$, where $|X| \geq 4$, is *diversified* if there are no distinct four elements $x, y, z, w \in X$ and $\lambda, \mu, \theta \in \mathbb{R}$ with $\lambda+\mu+\theta = 1$ such that $v(x, \cdot) \leq \lambda v(y, \cdot)) + \mu v(z, \cdot) + \theta v(w, \cdot)$. If $|X| < 4$, v is diversified if no row in v is dominated by an affine combination of the others.

We can finally state

Theorem 1 *: Let there be given X, $\mathbb{C}$, and $\{\succsim_M\}_{M\in\mathbb{M}}$ satisfying the richness assumption as above. Then the following two statements are equivalent:*

(i) $\{\succsim_M\}_{M\in\mathbb{M}}$ *satisfy A1-A4;*

(ii) There is a diversified matrix $v : X \times \mathbb{C} \to \mathbb{R}$ that respects case equivalence and such that:

$$(*) \quad \begin{cases} \text{for every } M \in \mathbb{M} \text{ and every } x, y \in X, \\ \\ x \succsim_M y \quad \text{iff} \quad \sum_{c\in M} v(x, c) \geq \sum_{c\in M} v(y, c) \ , \end{cases}$$

Furthermore*, in this case the matrix v is unique in the following sense: v and u both satisfy $(*)$ and respect case equivalence iff there are a scalar $\lambda > 0$ and a matrix $\beta : X \times \mathbb{C} \to \mathbb{R}$ with identical rows (i.e., with constant columns), that respects case equivalence, such that $u = \lambda v + \beta$.*

Observe that, by the richness assumption, $\mathbb{C}$ is infinite, and therefore the matrix v has infinitely many columns. Moreover, the theorem does not restrict the cardinality of X, and thus v may also have infinitely many rows.

2.4 Notes on the proof

The Result is part of Theorem 1, and was stated only for expository purposes. We therefore prove only Theorem 1.

The notion of case equivalence allows us to reduce the discussion to vectors of non-negative integers. We define the set of *types* of cases to be the $\sim$-equivalence classes: $\mathbb{T} = \mathbb{C}/\sim$. Assume, for simplicity, that there are finitely many types and finitely many eventualities. Rather than referring to sets of specific cases (memories M), we focus on vectors of non-negative integers. Such a vector $I : \mathbb{T} \to \mathbb{Z}_+$ represents many equivalent memories by counting how many cases of each type are in each of these memories. Thus, instead of dealing with subsets of the set $\mathbb{C}$, most of the discussion will be conducted in the space $\mathbb{Z}_+^{\mathbb{T}}$. Next, using the combination axiom, we extend the family rankings $\{\succeq_I\}$ from $I \in \mathbb{Z}_+^{\mathbb{T}}$ to $I \in \mathbb{Q}_+^{\mathbb{T}}$.

Focusing on two eventualities, x and y, we divide the vectors $I \in \mathbb{Q}_+^{\mathbb{T}}$ to those that render x more likely than y, and to those that induce the opposite ranking. Completeness and combination are the key axioms that allow us to invoke a separating hyperplane theorem. With the aid of the Archimedean axiom, one can prove that the separating hyperplane precisely characterizes the memories for which x is (strongly or weakly) more likely than y.

If one has only two eventualities, the proof is basically complete. Most of the work is in showing that the hyperplanes, which were obtained for each *pair* of eventualities, can be represented by a single matrix. More concretely, the separation theorem applied to a pair x, y yields a vector v^{xy}, unique up to multiplication by a positive constant, such that x is at least as likely as y given memory I iff $v^{xy} \cdot I \geq 0$. One now wishes to find a vector v^x for each eventuality x such that v^{xy} is a positive multiple of $(v^x - v^y)$ (simultaneously

for all x, y).

This can be done if and only if there is a selection of vectors $\{v^{xy}\}_{x,y}$ (where each is given only up to a multiplicative constant) such that $v^{xz} = v^{xy} + v^{yz}$ for every triple x, y, z. It turns out that, due to transitivity, this can be done for every triple x, y, z *separately*. The diversity axiom guarantees that this can also be done for sets of four eventualities, and the proof proceeds by induction.

The final two steps of the proof deal with extensions to infinitely many types and to infinitely many eventualities.

2.5 Mathematical comments

Given any real matrix of order $|X| \times |\mathbb{C}|$, one can define for every $M \in \mathbb{M}$ a weak order on X through $(*)$. It is easy to see that it will satisfy A1 and A2. If the matrix also respects case equivalence, A3 will also be satisfied. However, these conditions do not imply A4. For example, A4 will be violated if a row in the matrix dominates another row. Since A4 is not necessary for a representation by a matrix v via $(*)$ (even if it respects case equivalence), one may wonder whether it can be dropped. The answer is given by the following.

Proposition 2 *Axioms A1, A2, and A3 do not imply the existence of a matrix v that satisfies $(*)$.*

Some remarks on cardinality are in order. Axiom A4 can only hold if the set of types, $\mathbb{T} = \mathbb{C}/\sim$, is large enough relatively to X. For instance, if there are two distinct eventualities, the diversity axiom requires that there be at least two different types of cases. The following remark states that six types suffice for X to have the cardinality of the continuum.

Remark 3 *For any $\mathbb{T}$ such that $|\mathbb{T}| \geq 6$, there exists X with cardinality $\aleph$ and $\{\succsim_M\}_{M\in\mathbb{M}}$ that satisfy A1-4.*

Finally, one may wonder whether $(*)$ implies that v respects case equivalence. The negative answer is given below.

Remark 4 *Condition* $(*)$ *does not imply that* v *respects case equivalence.*

3 Axiomatization of Kernel Methods

3.1 Estimation of a density function

Assume that X is a continuous random variable taking values in $\mathbb{R}^m$. Having observed a finite sample $(x_i)_{i\leq n}$, one is asked to estimate the density function of X. Kernel estimation (see Akaike (1954), Rosenblatt (1956), Parzen (1962), and Silverman (1986) for a survey) suggests the following. Choose a (so-called "kernel") function $k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ with the following properties: (i) $k(x, y)$ is a non-increasing function of $\|x - y\|$; (ii) for every $x \in \mathbb{R}^m$, $\int_{\mathbb{R}^m} k(x, y)dy = 1$.[2] Given the sample $(x_i)_{i\leq n}$, estimate the density function by

$$f(y|x_1, ..., x_n) \equiv \tfrac{1}{n} \textstyle\sum_{i\leq n} k(x_i, y).$$

The estimated density function f quantifies likelihood differences between (small neighborhoods of) various points y. Let us now assume that, given the sample $x \equiv (x_i)_{i\leq n}$, the predictor is only asked to provide *qualitative* "at least as likely as" relations $\{\succsim_x\}$ on $\mathbb{R}^m$, where $y \succ_x z$ is interpreted to mean that (a small neighborhood of) y is more likely than (the corresponding neighborhood of) z given the sample x. Assume that the order of observations in $x = (x_i)_{i\leq n}$ does not affect the ranking $\succsim_x$. In this case a vector x can be identified with a finite set of observations, each of which is an element of $\mathbb{R}^m$. To allow repetitions, one defines $\mathbb{C} = \mathbb{R}^m \times \mathbb{N}$, where a case $(z, j) \in \mathbb{R}^m \times \mathbb{N}$ is interpreted as the j-th appearance of z in $x = (x_i)_{i\leq n}$. For any

[2] More generally, the kernel may be a function of transformed coordinates. The following discussion does not depend on assumptions (i) and (ii) and they are retained merely for concreteness.

finite $M \subset \mathbb{C}$, $\succsim_M$ is defined as follows: $\succsim_M = \succsim_x$ for some $x = (x_i)_{i \leq n}$ such that, for every $z \in \mathbb{R}^m$, $\#\{i \leq n \,|\, x_i = z\} = \#\{(z,j) \,|\, (z,j) \in M\}$.

Our axioms appear to be rather plausible in this set-up. Thus, the theorem suggests that there are numbers, $v(y,x)$, such that $\succsim_x$ is represented by $g_x(y) = \sum_{i \leq n} v(y, x_i)$. Up to normalization by a factor of n, the function v can serve the role of the kernel function. That is, setting $k(x,y) = k(y,x) = nv(y,x)$, the function g_x (defined via v) coincides with the function f (defined via k).

One may impose additional conditions on the collection of relations $\{\succsim_x\}$ such that the derived function v is (i) bounded; (ii) decreasing in distance; (iii) continuous in x for every y. Such axioms would employ the special structure of cases in this application. For instance, one may require that, if $\|x_i - y\| = \|x_i' - y'\|$ and $\|x_i - z\| = \|x_i' - z'\|$ for every $i \leq n$, then $y \succsim_x z$ iff $y' \succsim_x z'$. Similarly, the natural topology on $\mathbb{R}^m$ can be used to state that the set of observations $x \in (\mathbb{R}^m)^n$, for which $y \succ_x z$, is open for all $y, z \in \mathbb{R}^m$.

Deriving specific results on kernel functions is beyond the scope of this paper. At this point, we wish to point out that our theorem can be interpreted as a normative justification of kernel estimation, as well as a way to calibrate the appropriate kernel function based on intuitive likelihood judgments. Importantly, the kernel function k (or v) in our model is derived from these qualitative judgments, rather than assumed primitive.

3.2 Kernel classification

Kernel methods are also used for classification problems. Assume that a classifier is confronted with a data point $y \in \mathbb{R}^m$, and it is asked to guess to which member of a finite set A it belongs. The classifier is equipped with a set of examples M. Each example is a data point $x \in \mathbb{R}^m$, with a known classification in A. Kernel classification methods would adopt a kernel function a above, and, given the point y, would guess that y belongs to a class $a \in A$ that maximizes the sum of $k(x,y)$ over all x's in memory that

were classified as a.

Our general framework can accommodate classification problems as well. The special feature of classification problems is that each past cases specifies one of the objects to be ranked, namely, the classes, interpreted as the "correct" answer. To allow repetitions, we define the set of cases $\mathbb{C}$ to be $\mathbb{R}^m \times A \times \mathbb{N}$. Thus, an example (x, a) that appears in memory twice will be coded as a pair of cases, $(x, a, 1)$ and $(x, a, 2)$, in each of which a data point x in $\mathbb{R}^m$ was encountered, and its correct classification is known to have been $a \in A$.

For the purpose of axiomatization, we assume that, for each data point y, the classifier can rank all possible classes, given every possible memory.[3] That is, for a given (finite) memory $M \subset \mathbb{R}^m \times A \times \mathbb{N}$, classes in A are ranked by $\succsim_{M,y} \subset A \times A$. We will assume that $(x, a, i) \sim (x, a, j)$ for all $x \in \mathbb{R}^m$, $a \in A$, and $i, j \in \mathbb{N}$. Our main result suggests that, if $\{\succsim_{M,y}\}_M$ satisfy A1-4, then there is a matrix $v \equiv v_y : A \times (\mathbb{R}^m \times A \times \mathbb{N}) \to \mathbb{R}$, respecting case equivalence, such that $\succsim_{M,y}$ is represented by v_y as follows:

$$a \succsim_{M,y} b \text{ iff } \sum_{(x,d,i)\in M} v_y(a, (x, d, i)) \geq \sum_{(x,d,i)\in M} v_y(b, (x, d, i)).$$

This formulation allows an example of class $d \notin \{a, b\}$ to affect the ranking of class a vs. class b for a given y. There may be situations where one may wish to allow such linkages. For instance, in classification of visual inputs into the Latin alphabet, it is possible that similar data points that are known to represent lower case "o" may make upper case "O" more likely than, say, upper case "A". Yet, one may also impose an additional specificity axiom, stating that the ranking between classes a and b depends solely on examples of a or of b. For example, one may state the following axiom:

[3] Admittedly, a standard classification problem is less demanding in that it only requires a choice of one class, and not a complete ordering of all classes. See the discussion of scoring rules in voting theory in Section 6 below, and the comparison of our result to those of Young (1975) and Myerson (1995).

A5 Specificity: For every $y, x \in \mathbb{R}^m$, $M \in \mathbb{M}$, distinct $a, b, d \in A$, and $i \in \mathbb{N}$, $a \succsim_{M,y} b$ iff $a \succsim_{M \cup \{(x,d,i)\},y} b$.

This axiom yields a more compact representation:

Proposition 5 *Assume that $\{\succsim_{M,y}\}_{M \in \mathbb{M}}$ satisfy A1-4 and thus have a representation as in Theorem 1 by a matrix v_y. Then A5 is satisfied iff there exists a function $s_y : \mathbb{R}^m \times A \to R$ such that*

$$v_y(a, (x, b, i)) = s_y(x, a) 1_{a=b}$$

(where $1_{a=b}$ is 1 if $a = b$ and zero otherwise.)

***Moreover**, in this case the function s_y is unique up to multiplication by a positive constant.*

This formulation is very similar to kernel classification methods. One obvious difference is that the function s_y above may depend not only on past examples x, but also on the class considered a. This flexibility might be useful. For instance, in the example above, one may wish to give the letter "e" precedence, as a potential classification, over the letter "o". However, should one wish to rule out this possibility, one may impose an additional symmetry axiom on classes. It states that a permutation of the classes in the examples results in the corresponding permutation in the ranking of these classes for the new data point. Formally, let $\pi : A \to A$ be a permutation. For $M \in \mathbb{M}$, define $\pi(M) \in \mathbb{M}$ by $(x, a, i) \in M$ iff $(x, \pi(a), i) \in \pi(M)$ for every $(x, a, i) \in \mathbb{C}$. Using this notation, we write

A6 Symmetry: For every permutation π, every $M \in \mathbb{M}$, and every $a, b \in A$, $a \succsim_M b$, iff $\pi(a) \succsim_{\pi(M)} \pi(b)$.

Proposition 6 *Assume that $\{\succsim_{M,y}\}_{M \in \mathbb{M}}$ satisfy A1-4 and thus have a representation as in Theorem 1 by a matrix v_y. Then A5 and A6 are satisfied iff there exists a function $s_y : \mathbb{R}^m \to R$ such that*

$$v_y(a, (x, b, i)) = s_y(x) 1_{a=b}.$$

In particular, under A1-A6,

$$for\ every\ M \in \mathbb{M}\ and\ every\ a, b \in A,$$

$$a \succeq_{M,y} b\ iff\ \sum_{(x,a,i)\in M} s_y(x) \geq \sum_{(x,b,i)\in M} s_y(x)$$

***Moreover**, in this case the function s_y is unique up to multiplication by a positive constant.*

It is only left to define $k(x, y) = s_y(x)$ to obtain the familiar kernel formulation. Observe, however, that the kernel function is given up to a *separate* multiplicative constant for every y. In particular, k need not be symmetric. It can be shown (see Gilboa and Schmeidler (1995)) that $k(\cdot, y)$ may be re-scaled (separately for each y) to become symmetric if, for every $x, y, z \in \mathbb{R}^m$,

$$k(x, y)k(y, z)k(z, x) = k(x, z)k(z, y)k(y, x).$$

As above, this axiomatization can be viewed as a normative justification of kernel methods, and also as a way to elicit the "appropriate" kernel function from qualitative ranking data. Again, our approach does not assume that a kernel function is given, but derives such a function together with the kernel classification rule.

3.3 Comparison with nearest-neighbor approaches

A popular alternative to kernel classification methods is offered by nearest neighbor methods. Given a new instance y, rather than ranging over all past cases of all classes, a (single) nearest neighbor approach suggests to find an example (x, a) in memory that minimizes a distance function $d(x, y)$, and to select a as the class to which y probably belongs.[4] This algorithm appears to be somewhat extreme: a single most-similar case belonging to class a may outweigh dozens of slightly less similar cases all belonging to class b.

[4] As in the case of kernel functions, we discuss the simplest version of nearest neighbors methods for simplicity.

Thus, the nearest neighbor methodology was generalized to use several (k) nearest neighbors (Fix and Hodges (1951, 1952)). In this approach, a simple majority vote among the k nearest neighbors generates the prediction for each case. Further, one may extend the decision rule to a weighted majority vote (Royall (1966)), in which, among the k nearest neighbors, those that are closer to the case at hand are weighted more heavily. One of the merits of the nearest neighbor approaches is that one can bound their asymptotic probability of error, relative to that of a Bayesian decision (Cover and Hart (1967)). Further, k-nearest neighbors rules in which k tends to infinity with the number of observations n but does so slowly ($k/n \to 0$) enjoy universal consistency: their probability of error tends to that of the Bayesian decision for any underlying distribution from which the data are drawn (Stone (1977); see Devroye, Gyorfi, and Lugosi (1996)).

As kernel classifiers, nearest neighbor classifiers are designed to select one possible class for a new data point, rather than to rank all possible classes. Yet it is straightforward to extend them to generate complete rankings: after choosing the most plausible class by a vote among the nearest neighbors, one can ignore all past cases of this class and use the remaining database to select among the other categories, and so forth.

Our axioms provide a new perspective for the comparison of kernel classifiers and nearest neighbor classifiers. First, we observe that nearest neighbor classifiers do not satisfy our Archimedean axiom. A single case that is closest to the problem at hand will outweigh less similar cases, no matter how many replications of these have been observed. Whether this is a merit or a flaw of the nearest neighbor approach is probably a matter of taste. It should be observed, however, that the Archimedean axiom is not very crucial to the core of our theory. Dropping this axiom, one may still use a separating theorem, with the modification that the separating hyperplane itself is not unambiguously categorized. Moreover, one may use lexicographic separation, or separation by non-standard numbers. Correspondingly, one

may obtain variations of Theorem 1 in which the numerical representation is lexicographic or employs non-standard numbers.

A more interesting perspective for comparison is suggested by the combination axiom. A nearest neighbor approach employing but a single neighbor satisfies this axiom. Indeed, a 1-nearest neighbor ranking conforms with the numerical representation of Theorem 1 provided that nonstandard numbers are allowed. By contrast, for $k > 1$ a majority vote among the k-nearest neighbors violates the combination axiom. Take, for concreteness, $k = 3$ (avoiding possible ties). Assume that a physician classifies a patient as "sick" or "healthy" according to the majority among the 3 closest known cases. Further, assume that for a given patient there are six relevant cases, denoted $\{a, b, c, d, e, f\}$. Cases a, d are of distance 1 from the case at hand, and involved healthy patients, while cases b, c, e, f are of distance 2, and involved sick patients. Given each of the databases $\{a, b, c\}$ and $\{d, e, f\}$, the physician would classify the patient as sick. Given their union, she would conclude that the patient is healthy. It appears that this violation of the combination axiom is counter-intuitive.

If k grows with the number of observations n, our physician may use more cases out of a database containing six cases than out of each of the two databases containing three. Yet, k-nearest neighbors with $k > 1$ are in fundamental conflict with the combination axiom. Intuitively, the reason is that in these approaches the weight assigned to a past case depends not only on the inherent similarity between it and the case at hand, but also on its relative ranking, among other past cases, in terms of similarity. By contrast, our approach gives each case a weight that depends only on itself, irrespective of how many or which other cases are being taken into account.

4 Axiomatization of Maximum Likelihood Ranking

So far, the main interpretation of our model deals with ranking eventualities in a specific prediction problem. But the model can also interpreted as referring to ranking of theories or hypotheses given a set of observations. Observe that the axioms we formulated apply to this case as well. In particular, our main requirements are that theories be ranked by a weak order for every memory, and that, if theory x is more plausible than theory y given each of two disjoint memories, x should also be more plausible than y given the union of these memories.

Assume, therefore, that Theorem 1 holds. Suppose that, for each case c, $v(x,c)$ is bounded from above, and choose a representation v where $v(x,c) < 0$ for every theory x and case c. Define $p(c|x) = \exp(v(x,c))$, so that $\log(p(c|x)) = v(x,c)$. Our result states that, for every two theories x, y:

$$(*) \qquad x \succsim_M y \qquad \text{iff} \qquad \textstyle\sum_{c\in M} v(x,c) \geq \sum_{c\in M} v(y,c),$$

which is equivalent to

$$\exp\left(\textstyle\sum_{c\in M} v(x,c)\right) \geq \exp\left(\sum_{c\in M} v(y,c)\right)$$

or

$$\textstyle\prod_{c\in M} p(c|x) \geq \prod_{c\in M} p(c|y)$$

In other words, should a predictor rank theories in accordance with A1-A4, there exist conditional probabilities $p(c|x)$, for every case c and theory x, such that the predictor ranks theories as if by their likelihood functions, under the implicit assumption that the cases were stochastically independent. On the one hand, this result can be viewed as a normative justification of the likelihood rule: any method of ranking theories that is not equivalent to ranking by likelihood (for *some* conditional probabilities $p(c|x)$) has to violate one of our axioms. On the other hand, our result can be descriptively

interpreted, saying that likelihood rankings of theories are rather prevalent. One need not consciously assign conditional probabilities $p(c|x)$ for every case c given every theory x, and one need not know probability calculus in order to generate predictions in accordance with the likelihood criterion. Rather, whenever one satisfies our axioms, one may be ascribed conditional probabilities $p(c|x)$ such that one's predictions are in accordance with the resulting likelihood functions. Thus, relatively mild consistency requirements imply that one predicts *as if* by likelihood functions.

Finally, our result may be used to elicit the subjective conditional probabilities $p(c|x)$ of a predictor, given her qualitative rankings of theories. However, our uniqueness result is somewhat limited. In particular, for every case c one may choose a positive constant β_c and multiply $p(c|x)$ by β_c for all theories x, resulting in the same likelihood rankings. Similarly, one may choose a positive number α and raise all probabilities $\{p(c|x)\}_{c,x}$ to the power of α, again without changing the observed ranking of theories given possible memories. Thus there will generally be more than one set of conditional probabilities $\{p(c|x)\}_{c,x}$ that are consistent with $\{\succsim_M\}_{M\in\mathbb{M}}$.

The likelihood function relies on independence across cases. Conceptually, stochastic independence follows from two assumptions in our model. First, we have defined $\{\succsim_M\}_{M\in\mathbb{M}}$ where each M is a set. This implicitly assumes that only the number of repetitions of cases, and not their order, matters. This structural assumption is reminiscent of de Finetti's exchangeability condition (though the latter is defined in a more elaborate probabilistic model). Second, our combination axiom also has a flavor of independence. In particular, it rules out situations in which past occurrences of a case make future occurrences of the same case less likely.

5 Discussion of the Axioms

We argue that our axioms are generally rather plausible and that the prediction rule that they axiomatize is reasonably intuitive. The fact that this rule generalizes rankings by empirical frequencies may also serve as an argument in its favor. Moreover, it turns out that our key axioms are satisfied by well-known methods for various problems of learning, prediction, or inference, as shown in the previous two sections. This fact can also be cited as a piece of evidence that the axioms are indeed plausible.

But there are applications in which the axioms do not appear compelling. We discuss here several examples, trying to delineate the scope of applicability of the axioms, and to identify certain classes of situations in which they may not apply. In these situations one should take the linear aggregation rule with a grain of salt, for descriptive and for normative purposes alike.

In the following discussion we do not dwell on the first axiom, namely, that likelihood rankings are weak orders. This axiom and its limitations have been extensively discussed in decision theory, and there seems to be no special arguments for or against it in our specific context. We also have little to add to the discussion of the diversity axiom. While it does not appear to pose conceptual difficulties, there are no fundamental reasons to insist on its plausibility either. One may well be interested in other assumptions that would allow a representation as in $(*)$ by a matrix v that is not necessarily diversified. We therefore focus on the combination and the Archimedean assumptions.

Mis-specified cases Consider a cat, say Lucifer, who every so often dies and then may or may not resurrect. Suppose that, throughout history, many other cats have been observed to resurrect exactly eight times. If Lucifer had died and resurrected four times, and now died for the fifth time, we'd expect him to resurrect again. But if we double the number of cases, implying that we are now observing the ninth death, we would not expect Lucifer to be with us again. Thus, one may argue, the combination axiom does not seem

to be very compelling.

Obviously, this example assumes that all of Lucifer's deaths are equivalent. While this may be a reasonable assumption of a naive observer, the cat connoisseur will be careful enough to distinguish "first death" from "second death", and so forth. Thus, this example suggests that one has to be careful in the definition of a "case" (and of case equivalence) before applying the combination axiom.

Mis-specified theories Suppose that one wishes to determine whether a coin is biased. A memory with 1,000 repetitions of "Head", as well as a memory with 1,000 repetitions of "Tail" both suggest that the coin is indeed biased, while their union suggests that it is not. As mentioned above, our combination axiom makes an implicit assumption of stochastic independence. Under this assumption, it is highly unlikely to observe such memories. That is, adopting the combination axiom entails an implicit assumption that such anomalies will not occur. But this example also shows that ambiguity in the formulation of theories may make us reject the combination axiom. Specifically, the theory that the coin is "biased", without specifying in what way it biased, may be viewed as a mis-specified theory.

Theories about patterns A related class of examples deal with concepts that describe, or are defined by patterns, sequences, or sets of cases. Assume that a single case consists of 100 tosses of a coin. A complex sequence of 100 tosses may lend support to the hypothesis that the coin generates random sequences. But many repetitions of the very same sequence would undermine this hypothesis. Observe that "the coin generates random sequences" is a statement about *sequences* of cases. Similarly, statements such as "The weather always surprises" or "History repeats itself" are about sequences of cases, and are therefore likely to generate violations of the combination axiom.

Overwhelming evidence There are situations in which a single case may outweigh any number of repetitions of other cases, in contradiction to the

Archimedean axiom. For instance, a physician may find a single observation, taken from the patient she is currently treating, more relevant than any number of observations taken from other patients.[5] In the context of ranking theories, it is possible that a single case c constitutes a direct refutation of a theory x. If another theory y was not refuted by any case in memory, a single occurrence of case c will render theory x less plausible than theory y regardless of the number of occurrences of other cases, even if these lend more support to x than to y.[6] In such a case, one would like to assign conditional probability of zero to case c given theory x, or, equivalently, to set $v(x, c)$ to $-\infty$. More generally, one may extend Theorem 1 to provide representations by non-standard numbers, allowing several levels of impossibility as well.

Second-order induction An important class of examples in which we should expect the combination axiom to be violated, for descriptive and normative purposes alike, involves learning of the similarity function. For instance, assume that one database contains but one case, in which Mary chose restaurant x over y.[7] One is asked to predict what John's decision would be. Having no other information, one is likely to assume some similarity of tastes between John and Mary and to find it more plausible that John would prefer x to y as well. Next assume that in a second database there are no observed choices (by anyone) between x and y. Hence, based on this database alone, it would appear equally likely that John would choose x as that he would y. Assume further that this database does contain many choices between other pairs of restaurants, and it turns out that John and Mary consistently choose different restaurants. When combining the two databases, it makes sense to predict that John would choose y over x.

This is an instance in which the similarity function is learned from cases. Linear aggregation of cases by fixed weights embodies learning *by* a similarity

[5]Indeed, the nearest neighbor approach to classification problems violates the Archimedean axiom.

[6]This example is due to Peyton Young.

[7]This is a variant of an example by Sujoy Mukerji.

function. But it does not describe how this function *itself* is learned. In Gilboa and Schmeidler (2001) we call this process "second-order induction" and argue that the linear formula should only be taken as a very rough approximation when such a process is involved.

Combinations of inductive and deductive reasoning Another important class of examples in which the combination axiom is not very reasonable consists of prediction problems in which some structure is given. Consider a simple regression problem where a variable x is used to predict another variable y. Does the method of ordinary least squares satisfy our axioms? The answer depends on the unit of analysis. If we consider the regression equation $y = a + bx$ and attempt to predict the values of a and b given a sample $\{(x_i, y_i)\}_{i \leq n}$, the answer is in the affirmative. The least squares estimators of the parameters a and b are maximum likelihood estimators in the standard statistical model of regression analysis. If we define $\succsim_M$ by the likelihood function for this model, the collection $\{\succsim_M\}_M$ will satisfy the combination axiom. But if the units of analysis are the particular values of y for a new value of x, the answer is negative.

The reason is that the regression model is structured enough to allow some deductive reasoning. In ranking the plausibility of values of y for a given value of x, one makes two steps. First, one uses inductive reasoning to obtain estimates of the parameters a and b. Then, espousing a belief in the linear model, one uses these estimates to rank values of y by their plausibility. This second step involves deductive reasoning, exploiting the particular structure of the model. While the combination axiom is rather plausible for the first, inductive step, there is no reason for it to hold also for the entire inductive-deductive process.[8]

To consider another example, assume that a coin is about to be tossed in an i.i.d. manner. The parameter of the coin is not known, but one knows

[8]One may also view the examples discussed above under "theories about patterns" and under "overwhelming evidence" as special cases of structured inference.

probability rules that allow one to infer likelihood rankings of outcomes given any value of the unknown parameter. Again, when one engages in inference about the unknown parameter, one performs only inductive reasoning, and the combination axiom seems plausible. But when one is asked about particular outcomes, one uses inductive reasoning as well as deductive reasoning. In these cases, the combination axiom is too crude.[9]

In conclusion, there are classes of counterexamples to our axioms that result from under-specification of cases, of eventualities, or of memories. There are others that are more fundamental. Among these, two seem to deserve special attention. First, there are situations where second-order induction is involved, and the similarity function itself is learned. Indeed, our model deals with accumulated evidence but does not capture the emergence of new insights. Second, there are problems where some theoretical structure is assumed, and it can be used for deductive inferences. Our model captures some forms of inductive reasoning, but does not provide a full account of inferential processes involving a combination of inductive and deductive reasoning.

6 Other Interpretations

Decisions Theorem 1 can also have other interpretations. In particular, the objects to be ranked may be possible acts, with the interpretation of ranking as preferences. In this case, $v(x, c)$ denotes the support that case c lends to the choice of act x. The decision rule that results generalizes most of the decision rules of case-based decision theory (Gilboa and Schmeidler (2001)), as well as expected utility maximization, if beliefs are generated from cases in an additive way (see Gilboa and Schmeidler (1999)). Gilboa, Schmeidler, and Wakker (1999) apply this theorem, as well as an alternative approach, to axiomatize a theory of case-based decisions in which both the similarity

[9]We have received several counterexamples to the combination axiom that are, in our view, of this nature. In particular, we would like to thank Bruno Jullien, Klaus Nehring, and Ariel Rubinstein.

function between problem-act pairs and the utility function of outcomes are derived from preferences. This model generalizes Gilboa and Schmeidler (1997), in which the utility function is assumed given and only the similarity function is derived from observed preferences.

Voting Another interpretation is the derivation of scoring rules in voting theory, where cases are replaced by ballots, and eventualities – by candidates. Scoring rules have been axiomatized by Smith (1973), Young (1975), and Myerson (1995). Whereas these models bear similarity to ours, several differences exist. First, Smith and Young restrict the set of ballots to the permutations of the set of candidates. Myerson allows an abstract set of ballots, but employs a neutrality axiom, which relates the set of candidates to the set of ballots. By contrast, our model does not presuppose any relationship between cases and eventualities. This allows very different interpretations as in Sections 3 and 4 above, as well as scoring rules that do not satisfy symmetry. Second, while Smith assumes that selection is a complete ordering of the candidates, Young and Myerson assume only a choice correspondence. Our model, like Smith's, therefore assumes that more information is given in the data. On the other hand, we derive an almost-unique matrix v and can thus claim to provide a definition of the scoring weights by in-principle observable qualitative plausibility rankings.

Expected utility One may also use our result to derive a utility function in a two-person game, or in a "game against nature", that is, in a decision problem. Assume that a decision matrix, or a two-person game, is given, where the outcomes are abstract entities. Suppose that, for each mixed strategy of nature, the decision maker has a ranking over her pure strategies. Should these preferences satisfy our axioms (with some obvious modifications), one may attach a number to each outcome in the matrix such that preferences are given by maximization of the expectation of these numbers. Of course, these "utility" numbers will be unique only up to additions of numbers to columns, and multiplication of the entire matrix by a positive

constant. Indeed, these are transformations that do not change the structure of the best response correspondence is a game. (See Gilboa and Schmeidler (1999) for more details and for comparison with the axiomatization of expected utility maximization by von Neumann and Morgenstern (1944).)

Probabilities The main contribution of Gilboa and Schmeidler (1999) is to generalize the scope of prediction from eventualities to events. That is, in that paper we assume that the objects to be ranked belong to an algebra of subsets of a given set. Additional assumptions are imposed so that similarity values are additive with respect to the union of disjoint sets. Further, it is shown that ranking by empirical frequencies can also be axiomatically characterized in this set-up. Finally, tying the derivation of probabilities with expected utility maximization, one obtains a characterization of subjective expected utility maximization in face of uncertainty. As opposed to the behavioral axiomatic derivations of de Finetti (1937) and Savage (1954), which infer beliefs from decisions, this axiomatic derivation follows a presumed cognitive path leading from belief to decision.

Appendix: Proofs

Proof of Observation:

It is obvious that $\sim$ is reflexive and symmetric. To show that it is transitive, assume that $c \sim d$ and $d \sim e$ for distinct c, d, e. Let M be such that $c, e \notin M$. If $d \notin M$, then $\succsim_{M\cup\{c\}} = \succsim_{M\cup\{d\}}$ by $c \sim d$ and $\succsim_{M\cup\{d\}} = \succsim_{M\cup\{c\}}$ by $d \sim e$, and $\succsim_{M\cup\{c\}} = \succsim_{M\cup\{e\}}$ follows. If $d \in M$, define $N = M\backslash\{d\}$. Since $c, d \notin N \cup \{e\}$, $c \sim d$ implies $\succsim_{N\cup\{e\}\cup\{c\}} = \succsim_{N\cup\{e\}\cup\{d\}}$. Similarly, since $d, e \notin N\cup\{c\}$, $d \sim e$ implies $\succsim_{N\cup\{c\}\cup\{d\}} = \succsim_{N\cup\{c\}\cup\{e\}}$. It follows that $\succsim_{M\cup\{c\}} = \succsim_{N\cup\{c,d\}} = \succsim_{N\cup\{c,e\}} = \succsim_{N\cup\{d,e\}} = \succsim_{M\cup\{e\}}$.□

Proof of Theorem 1:

Let $\mathbb{T} = \mathbb{C}/\sim$ be the set of *types* of cases.[10] We prove the theorem in three steps. First we assume that there are finitely many types, that is, that

[10] $\mathbb{C}/\sim$ is the set of equivalence classes of $\sim$.

$|\mathbb{T}| < \infty$. In this case the proof relies on an auxiliary result that is of interest in its own right. Since the proof of this theorem applies to an infinite set of eventualities X, we do not restrict the cardinality of X in this case. Step 2 proceeds to deal with the case in which $|\mathbb{T}|$ is unrestricted, but X is finite. Lastly, Step 3 deals with the general case in which both $|X|$ and $|\mathbb{T}|$ are unrestricted.

In all three steps, memories in $\mathbb{M}$ are represented by vectors of non-negative integers, counting how many cases of each type appear in memory. Formally, for every $T \subset \mathbb{T}$ define $\mathbb{J}_T = \mathbb{Z}_+^T = \{I \,|\, I : T \to \mathbb{Z}_+\}$ where $\mathbb{Z}_+$ stands for the non-negative integers. $I \in \mathbb{J}_T$ is interpreted as a counter vector, where $I(t)$ counts how many cases of type t appear in the memory represented by I. For $I \in \mathbb{J}_T$, if $\{t \,|\, I(t) > 0\}$ is finite, define $\succsim_I \subset X \times X$ as follows. Choose $M \in \mathbb{M}$ such that $M \subset \cup_{t \in T} t$ (recall that $t \subset \mathbb{C}$ is an equivalence class of cases) and $I(t) = \#(M \cap t)$ for all $t \in T$, and define $\succsim_I = \succsim_M$. Such a set M exists since, by the richness assumption, $|t| \geq \aleph_0$ for all $t \in \mathbb{T}$. For this reason, such a set M is not unique. However, if both $M_1, M_2 \in \mathbb{M}$ satisfy these properties, then $M_1 \sim M_2$ and $\succsim_{M_1} = \succsim_{M_2}$. Hence $\succsim_I$ is well-defined.

Moreover, this definition implies the following property, which will prove useful in the sequel: if $I \in \mathbb{J}_T$ and $I' \in \mathbb{J}_{T'}$ where $T \subset T'$, $I'(t) = I(t)$ for $t \in T$ and $I'(t) = 0$ for $t \in T' \backslash T$, then $\succsim_I = \succsim_{I'}$. Another obvious observation, to be used later, is that for every $M \in \mathbb{M}$ there exist a finite $T \subset \mathbb{T}$ and $I \in \mathbb{J}_T$ such that $M \subset \cup_{t \in T} t$ and $I(t) = \#(M \cap t)$ for all $t \in T$.

Step 1: The case $|\mathbb{T}| < \infty$.

Denote the set of all counter vectors by $\mathbb{J} = \mathbb{J}_{\mathbb{T}} = \mathbb{Z}_+^{\mathbb{T}}$. For $I \in \mathbb{J}$, define $\succsim_I \subset X \times X$ as above. We now re-state the main theorem for this case, in the language of counter vectors. In the following, algebraic operations on $\mathbb{J}$ are performed pointwise.

A1* Order: For every $I \in \mathbb{J}$, $\succsim_I$ is complete and transitive on X.

A2* Combination: For every $I, J \in \mathbb{J}$ and every $x, y \in X$, if $x \succsim_I y$

$(x \succ_I y)$ and $x \succsim_J y$, then $x \succsim_{I+J} y$ $(x \succ_{I+J} y)$.

A3* Archimedean Axiom: For every $I, J \in \mathbb{J}$ and every $x, y \in X$, if $x \succ_I y$, then there exists $l \in N$ such that $x \succ_{lI+J} y$.

Observe that in the presence of Axiom 2, Axiom 3 also implies that for every $I, J \in \mathbb{J}$ and every $x, y \in X$, if $x \succ_I y$, then there exists $l \in \mathbb{N}$ such that for all $k \geq l$, $x \succ_{kI+J} y$.

A4* Diversity: For every list (x, y, z, w) of distinct elements of X there exists $I \in \mathbb{J}$ such that $x \succ_I y \succ_I z \succ_I w$. If $|X| < 4$, then for any strict ordering of the elements of X there exists $I \in \mathbb{J}$ such that $\succ_I$ is that ordering.

Theorem 7 *: Let there be given X, $\mathbb{T}$, and $\{\succsim_I\}_{I \in \mathbb{J}}$ as above. Then the following two statements are equivalent:*

(i) $\{\succsim_I\}_{I \in \mathbb{J}}$ satisfy A1-A4*;*

(ii) There is a diversified matrix $v : X \times \mathbb{T} \to \mathbb{R}$ such that:

$$(**) \quad \begin{cases} \text{for every } I \in \mathbb{J} \text{ and every } x, y \in X, \\ x \succsim_I y \quad \text{iff} \quad \sum_{t \in \mathbb{T}} I(t) v(x,t) \geq \sum_{t \in \mathbb{T}} I(t) v(y,t) \ , \end{cases}$$

***Furthermore**, in this case the matrix v is unique in the following sense: v and u both satisfy $(**)$ iff there are a scalar $\lambda > 0$ and a matrix $\beta : X \times \mathbb{T} \to \mathbb{R}$ with identical rows (i.e., with constant columns) such that $u = \lambda v + \beta$.*

Theorem 7 is reminiscent of the main result in Gilboa and Schmeidler (1997). In that work, cases are assumed to involve numerical payoffs, and algebraic and topological axioms are formulated in the payoff space. Here, by contrast, cases are not assumed to have any structure, and the algebraic and topological structures are given by the number of repetitions. This fact introduces two main difficulties. First, the space of "contexts" for which preferences are defined is not a Euclidean space, but only integer points thereof. This requires some care with the application of separation theorems.

Second, repetitions can only be non-negative. This fact introduces several complications, and, in particular, changes the algebraic implication of the diversity condition.

Before proceeding with the proof, we find it useful to present a condition that is equivalent to diversification of a matrix. We will use it both for the matrix $v : X \times \mathbb{T} \to R$ of Theorem 7 and the matrix $v : X \times \mathbb{C} \to R$ of Theorem 1. We therefore state it for an abstract set of columns:

Proposition 8 *Let Y be a set. Assume first $|X| \geq 4$. A matrix $v : X \times Y \to R$ is diversified iff for every list (x, y, z, w) of distinct elements of X, the convex hull of differences of the row-vectors $(v(x, \cdot) - v(y, \cdot))$, $(v(y, \cdot) - v(z, \cdot))$, and $(v(z, \cdot) - v(w, \cdot))$ does not intersect $\mathbb{R}^Y_-$. Similar equivalence holds for the case $|X| < 4$.*

Proof: We prove the lemma for the case $|X| \geq 4$. The proof for $|X| < 4$ is similar. Assume first that a matrix v is diversified. Assume that the conclusion does not hold. Hence, there are distinct $x, y, z, w \in X$ and $\alpha, \beta, \gamma \geq 0$ with $\alpha + \beta + \gamma = 1$ such that

$$\alpha(v(x, \cdot) - v(y, \cdot)) + \beta(v(y, \cdot) - v(z, \cdot)) + \gamma(v(z, \cdot) - v(w, \cdot)) \leq 0.$$

If $\alpha > 0$, then

$$v(x, \cdot) \leq \tfrac{\alpha - \beta}{\alpha} v(y, \cdot) + \tfrac{\beta - \gamma}{\alpha} v(z, \cdot) + \tfrac{\gamma}{\alpha} v(w, \cdot)$$

which means that $v(x, \cdot)$ is dominated by an affine combination of $\{v(y, \cdot), v(z, \cdot), v(w, \cdot)\}$, in contradiction to the fact that v is diversified. If $\alpha = 0$, then, by a similar argument, if $\beta > 0$, then $v(y, \cdot)$ is dominated by an affine combination of $\{v(z, \cdot), v(w, \cdot)\}$. Finally, if $\alpha = \beta = 0$, then $v(z, \cdot)$ is dominated by $v(w, \cdot)$.

For the converse direction, assume that the convex hull of $\{(v(x, \cdot) - v(y, \cdot)), (v(y, \cdot) - v(z, \cdot)), (v(z, \cdot) - v(w, \cdot))\}$ (over all lists (x, y, z, w) of distinct elements in X) does not intersect $\mathbb{R}^Y_-$ but that, contrary to diversity of v, there are distinct $x, y, z, w \in X$ and $\lambda, \mu, \theta \in \mathbb{R}$ with $\lambda + \mu + \theta = 1$ such that

$$(+) \qquad v(x,\cdot) \leq \lambda v(y,\cdot) + \mu v(z,\cdot) + \theta v(w,\cdot).$$

Since $\lambda + \mu + \theta = 1$, at least one of λ, μ, θ is non-negative. Assume, w.l.o.g., that $\theta \geq 0$. Hence $\lambda + \mu = 1 - \theta \leq 1$. This means that at least one of λ, μ cannot exceed 1. Assume, w.l.o.g., that $\lambda \leq 1$. Inequality $(+)$ can be written as

$$v(x,\cdot) - \lambda v(y,\cdot) - \mu v(z,\cdot) - \theta v(w,\cdot) \leq 0$$

or, equivalently,

$$(v(x,\cdot) - v(y,\cdot)) + (1-\lambda)(v(y,\cdot) - v(z,\cdot)) + (1-\lambda-\mu)(v(z,\cdot) - v(w,\cdot)) \leq 0.$$

Since $1 - \lambda \geq 0$ and $1 - \lambda - \mu = \theta \geq 0$, dividing by the sum of the coefficients yields a contradiction to the convex hull condition.□

Proof of Theorem 7: We present the proof for the case $|X| \geq 4$. The proofs for the cases $|X| = 2$ and $|X| = 3$ will be described as by-products along the way.

We start by proving that (i) implies (ii). We first note that the following homogeneity property holds:

Claim 1 *For every $I \in \mathbb{Z}_+^{\mathbb{T}}$ and every $k \in \mathbb{N}$, $\succsim_I = \succsim_{kI}$.*

Proof: Follows from consecutive application of the combination axiom. □

In view of this claim, we extend the definition of $\succsim_I$ to functions I whose values are non-negative rationals. Given $I \in \mathbb{Q}_+^{\mathbb{T}}$, let $k \in \mathbb{N}$ be such that $kI \in \mathbb{Z}_+^{\mathbb{T}}$ and define $\succsim_I = \succsim$. $\succsim_I$ is well-defined in view of Claim 1. By the definition and Claim 1 we also have:

Claim 2 ***(Homogeneity)*** *For every $I \in \mathbb{Q}_+^{\mathbb{T}}$ and every $q \in \mathbb{Q}$, $q > 0$: $\succsim_{qI} = \succsim_I$.*

Claim 2, A1*, and A2* imply:

Claim 3 *(The **order axiom**) For every $I \in \mathbb{Q}_+^{\mathbb{T}}$, $\succsim_I$ is complete and transitive on X, and (the **combination axiom**) for every $I, J \in \mathbb{Q}_+^{\mathbb{T}}$ and every $x, y \in X$ and $p, q \in \mathbb{Q}$, $p, q > 0$: if $x \succsim_I y$ $(x \succ_I y)$ and $x \succsim_J y$, then $x \succsim_{pI+qJ} y$ $(x \succ_{pI+qJ} y)$.*

Two special cases of the combination axiom are of interest: (i) $p = q = 1$, and (ii) $p + q = 1$. Claims 2 and 3, and the Archimedean axiom, A3*, imply the following version of the axiom for the $\mathbb{Q}_+^{\mathbb{T}}$ case:

Claim 4 *(The **Archimedean axiom**) For every $I, J \in \mathbb{Q}_+^{\mathbb{T}}$ and every $x, y \in X$, if $x \succ_I y$, then there exists $r \in [0, 1) \cap \mathbb{Q}$ such that $x \succ_{rI+(1-r)J} y$.*

It is easy to conclude from Claim 3 and 4 that for every $I, J \in \mathbb{Q}_+^{\mathbb{T}}$ and every $x, y \in X$, if $x \succ_I y$, then there exists $r \in [0, 1) \cap \mathbb{Q}$ such that $x \succ_{pI+(1-p)J} y$ for every $p \in (r, 1) \cap \mathbb{Q}$.

The following notation will be convenient for stating the first lemma. For every $x, y \in X$ let

$$A^{xy} \equiv \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid x \succ_I y\} \text{ and}$$
$$B^{xy} \equiv \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid x \succsim_I y\}.$$

Observe that by definition and A1*: $A^{xy} \subset B^{xy}$, $B^{xy} \cap A^{yx} = \emptyset$, and $B^{xy} \cup A^{yx} = \mathbb{Q}_+^{\mathbb{T}}$. The first main step in the proof of the theorem is:

Lemma 1 *For every distinct $x, y \in X$ there is a vector $v^{xy} \in \mathbb{R}^{\mathbb{T}}$ such that,*

(i) $B^{xy} = \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I \geq 0\}$;
(ii) $A^{xy} = \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I > 0\}$;
(iii) $B^{yx} = \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I \leq 0\}$;
(iv) $A^{yx} = \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I < 0\}$;
(v) Neither $v^{xy} \leq 0$ nor $v^{xy} \geq 0$;
(vi) $-v^{xy} = v^{yx}$.

Moreover, the vector v^{xy} satisfying (i)-(iv), is unique up to multiplication by a positive number.

The lemma states that we can associate with every pair of distinct eventualities $x, y \in X$ a separating hyperplane defined by $v^{xy} \cdot \xi = 0$ ($\xi \in \mathbb{R}^{\mathbb{T}}$), such that $x \succsim_I y$ iff I is in the half space defined by $v^{xy} \cdot I \geq 0$. Observe that if there are only two alternatives, Lemma 1 completes the proof of sufficiency: for instance, one may set $v^x = v^{xy}$ and $v^y = 0$. It then follows that $x \succsim_I y$ iff $v^{xy} \cdot I \geq 0$, i.e., iff $v^x \cdot I \geq v^y \cdot I$. More generally, we will show in the following lemmata that one can find a vector v^x for every alternative x, such that, for every $x, y \in X$, v^{xy} is a positive multiple of $(v^x - v^y)$.

Before starting the proof we introduce additional notation: let $\widehat{B}^{xy}$ and $\widehat{A}^{xy}$ denote the convex hulls (in $\mathbb{R}^{\mathbb{T}}$) of B^{xy} and A^{xy}, respectively. For a subset B of $\mathbb{R}^{\mathbb{T}}$ let $int(B)$ denote the set of interior points of B.

Proof of Lemma 1: We break the proof into several claims.

Claim 5 *For every distinct $x, y \in X$, $A^{xy} \cap int(\widehat{A}^{xy}) \neq \emptyset$.*

Proof: By the diversity axiom $A^{xy} \neq \emptyset$ for all $x, y \in X, x \neq y$. Let $I \in A^{xy} \cap \mathbb{Z}_+^{\mathbb{T}}$ and let $J \in \mathbb{Z}_+^{\mathbb{T}}$ with $J(t) > 1$ for all $t \in \mathbb{T}$. By the Archimedean axiom there is an $l \in \mathbb{N}$ such that $K = lI + J \in A^{xy}$. Let $(\xi_j)_{j=1}^{2^{|\mathbb{T}|}}$ be the $2^{|\mathbb{T}|}$ distinct vectors in $\mathbb{R}^{\mathbb{T}}$ with coordinates 1 and -1. For j, $(j = 1, ..., 2^{|\mathbb{T}|})$, define $\eta_j = K + \xi_j$. Obviously, $\eta_j \in \mathbb{Q}_+^{\mathbb{T}}$ for all j. By Claim 4 there is an $r_j \in [0, 1) \cap \mathbb{Q}$ such that $\varsigma_j = r_j K + (1 - r_j)\eta_j \in A^{xy}$ (for all j). Clearly, the convex hull of $\{ \varsigma_j \mid j = 1, ..., 2^{|\mathbb{T}|}\}$, which is included in $\widehat{A}^{xy}$, contains an open neighborhood of K. $\square$

Claim 6 *For every distinct $x, y \in X$, $\widehat{B}^{yx} \cap int(\widehat{A}^{xy}) = \emptyset$.*

Proof: Suppose, by way of negation, that for some $\xi \in int(\widehat{A}^{xy})$ there are $(\eta_i)_{i=1}^k$ and $(\lambda_i)_{i=1}^k$, $k \in \mathbb{N}$ such that for all i, $\eta_i \in B^{yx}$, $\lambda_i \in [0, 1]$, $\Sigma_{i=1}^k \lambda_i = 1$, and $\xi = \Sigma_{i=1}^k \lambda_i \eta_i$. Since $\xi \in int(\widehat{A}^{xy})$, there is a ball of radius $\varepsilon > 0$ around ξ included in $\widehat{A}^{xy}$. Let $\delta = \varepsilon/(2\Sigma_{i=1}^k ||\eta_i||)$ and for each i let $q_i \in \mathbb{Q} \cap [0, 1]$ such that $|q_i - \lambda_i| < \delta$, and $\Sigma_{i=1}^k q_i = 1$. Hence, $\eta = \Sigma_{i=1}^k q_i \eta_i \in \mathbb{Q}_+^{\mathbb{T}}$ and $||\eta - \xi|| < \varepsilon$, which, in turn, implies $\eta \in \widehat{A}^{xy} \cap \mathbb{Q}_+^{\mathbb{T}}$. Since for all $i : \eta_i \in B^{yx}$, consecutive

application of the combination axiom (Claim 3) yields $\eta = \Sigma_{i=1}^{k} q_i \eta_i \in B^{yx}$. On the other hand, η is a convex combination of points in $A^{xy} \subset \mathbb{Q}_+^{\mathbb{T}}$ and thus it has a representation with rational coefficients (because the rationals are an algebraic field). Applying Claims 3 consecutively as above, we conclude that $\eta \in A^{xy}$ – a contradiction.$\square$

The main step in the proof of Lemma 1: The last two claims imply that (for all $x, y \in X, x \neq y$) $\widehat{B}^{xy}$ and $\widehat{A}^{yx}$ satisfy the conditions of a separating hyperplane theorem. (Namely, these are convex sets, where the interior of one of them is non-empty and does not intersect the other set.) So there is a vector $v^{xy} \neq 0$ and a number c so that

$$v^{xy} \cdot I \geq c \quad \text{for every} \quad I \in \widehat{B}^{xy}$$
$$v^{xy} \cdot I \leq c \quad \text{for every } I \in \widehat{A}^{yx} \ .$$

Moreover,

$$v^{xy} \cdot I > c \quad \text{for every} \quad I \in int(\widehat{B}^{xy})$$
$$v^{xy} \cdot I < c \quad \text{for every } I \in int(\widehat{A}^{yx}) \ .$$

By homogeneity (Claim 2), $c = 0$. **Parts (i)-(iv) of the lemma** are restated as a claim and proved below.

Claim 7 *For all $x, y \in X, x \neq y$: $B^{xy} = \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I \geq 0\}$; $A^{xy} = \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I > 0\}$; $B^{yx} = \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I \leq 0\}$; and $A^{yx} = \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I < 0\}$.*

Proof: (a) $B^{xy} \subset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I \geq 0\}$ follows from the separation result and the fact that $z = 0$.

(b) $A^{xy} \subset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I > 0\}$: assume that $x \succ_I y$, and, by way of negation, $v^{xy} \cdot I \leq 0$. Choose a $J \in A^{yx} \cap int(\widehat{A}^{yx})$. Such a J exists by Claim 5. Since $z = 0$, J satisfies $v^{xy} \cdot J < 0$. By Claim 4 there exists $r \in [0, 1)$ such that $rI + (1-r)J \in A^{xy} \subset B^{xy}$. By (a), $v^{xy} \cdot (rI + (1-r)J) \geq 0$. But $v^{xy} \cdot I \leq 0$ and $v^{xy} \cdot J < 0$, a contradiction. Therefore, $A^{xy} \subset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I > 0\}$.

(c) $A^{yx} \subset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I < 0\}$: assume that $y \succ_I x$ and, by way of negation, $v^{xy} \cdot I \geq 0$. By Claim 5 there is a $J \in A^{xy}$ with $J \in int(\widehat{A}^{xy}) \subset int(\widehat{B}^{xy})$. The inclusion $J \in int(\widehat{B}^{xy})$ implies $v^{xy} \cdot J > 0$. Using the Archimedean axiom, there is an $r \in [0, 1)$ such that $rI + (1 - r)J \in A^{yx}$. The separation theorem implies that $v^{xy} \cdot (rI + (1-r)J) \leq 0$, which is impossible if $v^{xy} \cdot I \geq 0$ and $v^{xy} \cdot J > 0$. This contradiction proves that $A^{yx} \subset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I < 0\}$.

(d) $B^{yx} \subset \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I \leq 0\}$: assume that $y \succsim_I x$, and, by way of negation, $v^{xy} \cdot I > 0$. Let J satisfy $y \succ_J x$. By (c), $v^{xy} \cdot J < 0$. Define $r = (v^{xy} \cdot I)/(-v^{xy} \cdot J) > 0$. By homogeneity (Claim 2), $y \succ_{rJ} x$. By Claim 3, $I + rJ \in A^{yx}$. Hence, by (c), $v^{xy} \cdot (I + rJ) < 0$. However, direct computation yields $v^{xy} \cdot (I + rJ) = v^{xy} \cdot I + rv^{xy} \cdot J = 0$, a contradiction. It follows that $B^{yx} \subset \{I \in Q_+^{\mathbb{T}} \mid v^{xy} \cdot I \leq 0\}$.

(e) $B^{xy} \supset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I \geq 0\}$: follows from completeness and (c).

(f) $A^{xy} \supset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I > 0\}$: follows from completeness and (d).

(g) $A^{yx} \supset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I < 0\}$: follows from completeness and (a).

(h) $B^{yx} \supset \{I \in \mathbb{Q}_+^{\mathbb{T}} \mid v^{xy} \cdot I \leq 0\}$: follows from completeness and (b). □

Completion of the proof of the Lemma.

Part (v) of the Lemma, i.e., $v^{xy} \notin \mathbb{R}_+^{\mathbb{T}} \cup \mathbb{R}_-^{\mathbb{T}}$ for $x \neq y$, follows from the facts that $A^{xy} \neq \varnothing$ and $A^{yx} \neq \varnothing$. Before proving part (vi), we prove **uniqueness**.

Assume that both v^{xy} and u^{xy} satisfy (i)-(iv). In this case, $u^{xy} \cdot \xi \leq 0$ implies $v^{xy} \cdot \xi \leq 0$ for all $\xi \in \mathbb{R}_+^{\mathbb{T}}$. (Otherwise, there exists $I \in \mathbb{Q}_+^{\mathbb{T}}$ with $u^{xy} \cdot I \leq 0$ but $v^{xy} \cdot I > 0$, contradicting the fact that both v^{xy} and u^{xy} satisfy (i)-(iv).) Similarly, $u^{xy} \cdot \xi \geq 0$ implies $v^{xy} \cdot \xi \geq 0$. Applying the same argument for v^{xy} and u^{xy}, we conclude that $\{\xi \in \mathbb{R}_+^{\mathbb{T}} \mid v^{xy} \cdot \xi = 0\} = \{\xi \in \mathbb{R}_+^{\mathbb{T}} \mid u^{xy} \cdot \xi = 0\}$. Moreover, since $int(\widehat{A}^{xy}) \neq \varnothing$ and $int(\widehat{A}^{yx}) \neq \varnothing$, it follows that $\{\xi \in \mathbb{R}_+^{\mathbb{T}} \mid v^{xy} \cdot \xi = 0\} \cap int(\mathbb{R}_+^{\mathbb{T}}) \neq \varnothing$. This implies that $\{\xi \in \mathbb{R}^{\mathbb{T}} \mid v^{xy} \cdot \xi = 0\} = \{\xi \in \mathbb{R}^{\mathbb{T}} \mid u^{xy} \cdot \xi = 0\}$, i.e., that v^{xy} and u^{xy} have the same null set and are therefore a multiple of each other. That is, there exists α such that $u^{xy} = \alpha v^{xy}$. Since both satisfy (i)-(iv), $\alpha > 0$.

Finally, we prove **part (vi)**. Observe that both v^{xy} and $-v^{yx}$ satisfy (i)-(iv) (stated for the ordered pair (x, y)). By the uniqueness result, $-v^{xy} = \alpha v^{yx}$ for some positive number α. At this stage we redefine the vectors $\{v^{xy}\}_{x,y\in X}$ from the separation result as follows: for every unordered pair $\{x, y\} \subset X$ one of the two ordered pairs, say (y, x), is arbitrary chosen and then v^{xy} is rescaled such that $v^{xy} = -v^{yx}$. (If X is uncountable the axiom of choice has to be used.)□

Lemma 2 *For every three distinct eventualities, $x, y, z \in X$, and the corresponding vectors v^{xy}, v^{yz}, v^{xz} from Lemma 1, there are unique $\alpha, \beta > 0$ such that:*

$$\alpha v^{xy} + \beta v^{yz} = v^{xz} \ .$$

The key argument in the proof of Lemma 2 is that, if v^{xz} is not a linear combination of v^{xy} and v^{yz}, one may find a vector I for which $\succ_I$ is cyclical.

If there are only three alternatives $x, y, z \in X$, Lemma 2 allows us to complete the proof as follows: choose an arbitrary vector v^{xz} that separates between x and z. Then choose the multiples of v^{xy} and of v^{yz} defined by the lemma. Proceed to define $v^x = v^{xz}$, $v^y = \beta v^{yz}$, and $v^z = 0$. By construction, $(v^x - v^z)$ is (equal and therefore) proportional to v^{xz}, hence $x \succsim_I z$ iff $v^x \cdot I \geq v^z \cdot I$. Also, $(v^y - v^z)$ is proportional to v^{yz} and it follows that $y \succsim_I z$ iff $v^y \cdot I \geq v^z \cdot I$. The point is, however, that, by Lemma 2, we obtain the same result for the last pair: $(v^x - v^y) = (v^{xz} - \beta v^{yz}) = \alpha v^{xy}$ and $x \succsim_I y$ iff $v^x \cdot I \geq v^y \cdot I$ follows.

Proof of Lemma 2:

First note that for every three distinct eventualities, $x, y, z \in X$, if v^{xy} and v^{yz} are colinear, then for all I either $x \succ_I y \Leftrightarrow y \succ_I z$ or $x \succ_I y \Leftrightarrow z \succ_I y$. Both implications contradict diversity. Therefore any two vectors in $\{v^{xy}, v^{yz}, v^{xz}\}$ are linearly independent. This immediately implies the uniqueness claim of the lemma. Next we introduce

Claim 8 *For every distinct $x, y, z \in X$, and every $\lambda, \mu \in \mathbb{R}$, if $\lambda v^{xy} + \mu v^{yz} \leq 0$, then $\lambda = \mu = 0$.*

Proof: Observe that Lemma 1(v) implies that if one of the numbers λ, and μ is zero, so is the other. Next, suppose, per absurdum, that $\lambda\mu \neq 0$,and consider $\lambda v^{xy} \leq \mu v^{zy}$. If, say, $\lambda, \mu > 0$, then $v^{xy} \cdot I \geq 0$ necessitates $v^{zy} \cdot I \geq 0$. Hence there is no I for which $x \succ_I y \succ_I z$, in contradiction to the diversity axiom. Similarly, $\lambda > 0 > \mu$ precludes $x \succ_I z \succ_I y$; $\mu > 0 > \lambda$ precludes $y \succ_I x \succ_I z$; and $\lambda, \mu < 0$ implies that for no $I \in \mathbb{Q}_+^{\mathbb{T}}$ is it the case that $z \succ_I y \succ_I x$. Hence the diversity axioms holds only if $\lambda = \mu = 0$. $\square$

We now turn to the main part of the proof. Suppose that v^{xy}, v^{yz}, and v^{zx} are column vectors and consider the $|\mathbb{T}| \times 3$ matrix (v^{xy}, v^{yz}, v^{zx}) as a 2-person 0-sum game. If its value is positive, then there is an $\xi \in \Delta(\mathbb{T})$ such that $v^{xy} \cdot \xi > 0$, $v^{yz} \cdot \xi > 0$, and $v^{zx} \cdot \xi > 0$. Hence there is an $I \in \mathbb{Q}_+^{\mathbb{T}} \cap \Delta(\mathbb{T})$ that satisfies the same inequalities. This, in turn, implies that $x \succ_I y$, $y \succ_I z$, and $z \succ_I x$ - a contradiction.

Therefore the value of the game is zero or negative. In this case there are $\lambda, \mu, \zeta \geq 0$, such that $\lambda v^{xy} + \mu v^{yz} + \zeta v^{zx} \leq 0$ and $\lambda + \mu + \zeta = 1$. The claim above implies that if one of the numbers λ, μ and ζ is zero, so are the other two. Thus $\lambda, \mu, \zeta > 0$. We therefore conclude that there are $\alpha = \lambda/\zeta > 0$ and $\beta = \mu/\zeta > 0$ such that

$$(1) \qquad \alpha v^{xy} + \beta v^{yz} \leq v^{xz}$$

Applying the same reasoning to the triple z, y, and x, we conclude that there are $\gamma, \delta > 0$ such that

$$(2) \qquad \gamma v^{zy} + \delta v^{yx} \leq v^{zx}.$$

Summation yields

$$(3) \qquad (\alpha - \delta) v^{xy} + (\beta - \gamma) v^{yz} \leq 0.$$

Claim 8 applied to inequality (3) implies $\alpha = \delta$ and $\beta = \gamma$. Hence inequality (2) may be rewritten as $\alpha v^{xy} + \beta v^{yz} \leq v^{xz}$, which together with (1) yields the desired representation.□

Lemma 2 shows that, if there are more than three alternatives, the likelihood ranking of every triple of alternatives can be represented as in the theorem. The question that remains is whether these separate representations (for different triples) can be "patched" together in a consistent way.

Lemma 3 *There are vectors $\{v^{xy}\}_{x,y\in X,x\neq y}$, as in Lemma 1, such that for any three distinct acts, $x, y, z \in X$, the Jacobi identity $v^{xy} + v^{yz} = v^{xz}$ holds.*

Proof: The proof is by induction, which is transfinite if X is uncountably infinite. The main idea of the proof is the following. Assume that one has rescaled the vectors v^{xy} for all alternatives x, y in some subset of acts $A \subset X$, and one now wishes to add another act to this subset, $w \notin A$. Choose $x \in A$ and consider the vectors v^{xw}, v^{yw} for $x, y \in A$. By Lemma 2, there are unique positive coefficients α, β such that $v^{xy} = \alpha v^{xw} + \beta v^{wy}$. One would like to show that the coefficient $\alpha = \alpha_y$ does not depend on the choice of $y \in A$. We will show that, if α_y did depend on y, one would find that there are $x, y, z \in A$ such that the vectors v^{xw}, v^{yw}, v^{zw} are linearly dependent, and this would contradict the diversity axiom.

Claim 9 *Let $A \subset X$, $|A| \geq 3$, $w \in X\backslash A$. Suppose that there are vectors $\{v^{xy}\}_{x,y\in A,x\neq y}$, as in Lemma 1, and for any three distinct acts, $x, y, z \in X$, $v^{xy}+v^{yz} = v^{xz}$ holds. Then there are vectors $\{v^{xy}\}_{x,y\in A\cup\{w\},x\neq y}$, as in Lemma 1, and for any three distinct acts, $x, y, z \in X$, $v^{xy} + v^{yz} = v^{xz}$ holds.*

Proof: Choose distinct $x, y, z \in A$. Let $\hat{v}^{xw}$,$\hat{v}^{yw}$, and $\hat{v}^{zw}$ be the vectors provided by Lemma 1 when applied to the pairs (x, w), (y, w), and (z, w), respectively. Consider the triple $\{x, y, w\}$. By Lemma 2 there are unique coefficients $\lambda(\{x, w\}, y), \lambda(\{y, w\}, x) > 0$ such that

$$\text{(I) } v^{xy} = \lambda(\{x, w\}, y)\hat{v}^{xw} + \lambda(\{y, w\}, x)\hat{v}^{wy}$$

Applying the same reasoning to the triple $\{x, z, w\}$, we find that there are unique coefficients $\lambda(\{x, w\}, z), \lambda(\{z, w\}, x) > 0$ such that

$$v^{xz} = \lambda(\{x, w\}, z)\hat{v}^{xw} + \lambda(\{z, w\}, x)\hat{v}^{wz}.$$

or

$$\text{(II) } v^{zx} = \lambda(\{x, w\}, z)\hat{v}^{wx} + \lambda(\{z, w\}, x)\hat{v}^{zw}.$$

We wish to show that $\lambda(\{x, w\}, y) = \lambda(\{x, w\}, z)$. To see this, we consider also the triple $\{y, z, w\}$ and conclude that there are unique coefficients $\lambda(\{y, w\}, z), \lambda(\{z, w\}, y) > 0$ such that

$$\text{(III) } v^{yz} = \lambda(\{y, w\}, z)\hat{v}^{yw} + \lambda(\{z, w\}, y)\hat{v}^{wz}.$$

Since $x, y, z \in A$, we have

$$v^{xy} + v^{yz} + v^{zx} = 0$$

and it follows that the summation of the right-hand sides of (I), (II), and (III) also vanishes:

$$[\lambda(\{x, w\}, y) - \lambda(\{x, w\}, z)]\hat{v}^{xw} + [\lambda(\{y, w\}, z) - \lambda(\{y, w\}, x)]\hat{v}^{yw} + [\lambda(\{z, w\}, x) - \lambda(\{z, w\}, y)]\hat{v}^{zw} = 0.$$

If some of the coefficients above are not zero, the vectors $\{\hat{v}^{xw}, \hat{v}^{yw}, \hat{v}^{zw}\}$ are linearly dependent, and this contradicts the diversity axiom. For instance, if $\hat{v}^{xw}$ is a non-negative linear combination of $\hat{v}^{yw}$ and $\hat{v}^{zw}$, for no I will it be the case that $y \succ_I z \succ_I w \succ_I x$.

We therefore obtain $\lambda(\{x, w\}, y) = \lambda(\{x, w\}, z)$ for every $y, z \in A\backslash\{x\}$. Hence for every $x \in A$ there exists a unique $\lambda(\{x, w\}) > 0$ such that, for

every distinct $x, y \in A$ $v^{xy} = \lambda(\{x,w\})\hat{v}^{xw} + \lambda(\{y,w\})\hat{v}^{wy}$. Defining $v^{xw} = \lambda(\{x,w\})\hat{v}^{xw}$ completes the proof of the claim.□

To complete the proof of the lemma, we apply the claim consecutively. In case X is not countable, the induction is transfinite (and assumes that X can be well ordered).□

Note that Lemma 3, unlike Lemma 2, guarantees the possibility to rescale *simultaneously* all the v^{xy}-s from Lemma 1 such that the Jacobi identity will hold on X.

We now complete the proof that (i) implies (ii). Choose an arbitrary act, say, g in X. Define $v^g = 0$, and for any other alternative, x, define $v^x = v^{xg}$, where the v^{xg}-s are from Lemma 3.

Given $I \in \mathbb{Q}^{\mathbb{T}}_+$ and $x, y \in X$ we have:

$$x \succsim_I y \Leftrightarrow v^{xy} \cdot I \geq 0 \Leftrightarrow (v^{xg} + v^{gy}) \cdot I \geq 0 \Leftrightarrow$$
$$(v^{xg} - v^{yg}) \cdot I \geq 0 \Leftrightarrow v^x \cdot I - v^y \cdot I \geq 0 \Leftrightarrow v^x \cdot I \geq v^y \cdot I$$

The first implication follows from Lemma 1(i), the second from the Jacobi identity of Lemma 3, the third from Lemma 1(vi), and the fourth from the definition of the v^x-s. Hence, $(**)$ of the theorem has been proved.

It remains to be shown that the vectors defined above are such that $conv(\{v^x - v^y, v^y - v^z, v^z - v^w\}) \cap \mathbb{R}^{\mathbb{T}}_- = \emptyset$. Indeed, in Lemma 1(v) we have shown that $v^x - v^y \notin \mathbb{R}^{\mathbb{T}}_-$. To see this one only uses the diversity axiom for the pair $\{x, y\}$. Lemma 2 has shown, among other things, that a non-zero linear combination of $v^x - v^y$ and $v^y - v^z$ cannot be in $\mathbb{R}^{\mathbb{T}}_-$, using the diversity axiom for triples. Linear independence of all three vectors was established in Lemma 3. However, the full implication of the diversity condition will be clarified by the following lemma. Being a complete characterization, we will also use it in proving the converse implication, namely, that part (ii) of the theorem implies part (i). The proof of the lemma below depends on Lemma 1. It therefore holds under the assumptions that for any distinct $x, y \in X$ there is an I such that $x \succ_I y$.

Lemma 4 *For every list (x, y, z, w) of distinct elements of X, there exists $I \in \mathbb{J}$ such that*

$$x \succ_I y \succ_I z \succ_I w \quad \textit{iff} \quad conv(\{v^{xy}, v^{yz}, v^{zw}\}) \cap \mathbb{R}^{\mathbb{T}}_{-} = \emptyset \ .$$

Proof: There exists $I \in \mathbb{J}$ such that $x \succ_I y \succ_I z \succ_I w$ iff there exists $I \in \mathbb{J}$ such that $v^{xy} \cdot I, v^{yz} \cdot I, v^{zw} \cdot I > 0$. This is true iff there exists a probability vector $p \in \Delta(\mathbb{T})$ such that $v^{xy} \cdot p, v^{yz} \cdot p, v^{zw} \cdot p > 0$.

Suppose that v^{xy}, v^{yz}, and v^{zw} are column vectors and consider the $|\mathbb{T}| \times 3$ matrix (v^{xy}, v^{yz}, v^{zw}) as a 2-person 0-sum game. The argument above implies that there exists $I \in \mathbb{J}$ such that $x \succ_I y \succ_I z \succ_I w$ iff the maximin in this game is positive. This is equivalent to the minimax being positive, which means that for every mixed strategy of player 2 there exists $t \in \mathbb{T}$ that guarantees player 1 a positive payoff. In other words, there exists $I \in \mathbb{J}$ such that $x \succ_I y \succ_I z \succ_I w$ iff for every convex combination of $\{v^{xy}, v^{yz}, v^{zw}\}$ at least one entry is positive, i.e., $conv(\{v^{xy}, v^{yz}, v^{zw}\}) \cap \mathbb{R}^{\mathbb{T}}_{-} = \emptyset$. $\square$

This completes the proof that (i) implies (ii). $\square$

Part 2: (ii) implies (i)

It is straightforward to verify that if $\{\succsim_I\}_{i \in \mathbb{Q}^{\mathbb{T}}_{+}}$ are representable by $\{v^x\}_{x \in X}$ as in $(**)$, they have to satisfy Axioms 1-3. To show that Axiom 4 holds, we quote Lemma 4 of the previous part. $\square$

Part 3: Uniqueness

It is obvious that if $u^x = \alpha v^x + \beta$ for some scalar $\alpha > 0$, a vector $\beta \in \mathbb{R}^{\mathbb{T}}$, and all $x \in X$, then part (ii) of the theorem holds with the matrix u replacing v.

Suppose that $\{v^x\}_{x \in X}$ and $\{u^x\}_{x \in X}$ both satisfy $(**)$, and we wish to show that there are a scalar $\alpha > 0$ and a vector $\beta \in \mathbb{R}^{\mathbb{T}}$ such that for all $x \in X$, $u^x = \alpha v^x + \beta$. Recall that, for $x \neq y$, $v^x \neq \lambda v^y$ and $u^x \neq \lambda u^y$ for all $0 \neq \lambda \in \mathbb{R}$ by A4.

Choose $x \neq g$ ($x, g \in X$, g satisfies $v^g = 0$). From the uniqueness part of Lemma 1 there exists a unique $\alpha > 0$ such that $(u^x - u^g) = \alpha(v^x - v^g) = \alpha v^x$.

Define $\beta = u^g$.

We now wish to show that, for any $y \in X$, $u^y = \alpha v^y + \beta$. It holds for $y = g$ and $y = x$, hence assume that $x \neq y \neq g$. Again, from the uniqueness part of Lemma 1 there are unique $\gamma, \delta > 0$ such that

$$(u^y - u^x) = \gamma(v^y - v^x)$$
$$(u^g - u^y) = \delta(v^g - v^y) \ .$$

Summing up these two with $(u^x - u^g) = \alpha(v^x - v^g)$, we get

$$0 = \alpha(v^x - v^g) + \gamma(v^y - v^x) + \delta(v^g - v^y) = \alpha v^x + \gamma(v^y - v^x) - \delta v^y.$$

Thus

$$(\alpha - \gamma)v^x + (\gamma - \delta)v^y = 0 \ .$$

Since $v^x \neq v^g = 0$, $v^y \neq v^g = 0$, and $v^x \neq \lambda v^y$ if $0 \neq \lambda \in \mathbb{R}$, we get $\alpha = \gamma = \delta$. Plugging $\alpha = \gamma$ into $(u^y - u^x) = \gamma(v^y - v^x)$ proves that $u^y = \alpha v^y + \beta$. $\square$

This completes the proof of Theorem 7. $\square\square$

We now turn to complete the proof of Step 1. First we prove that (i) implies (ii). Assume that $\{\succcurlyeq_M\}_M$ satisfy A1-A4. It follows that $\{\succcurlyeq_I\}_I$ satisfy A1*-A4*. Therefore, there is a representation of $\{\succcurlyeq_I\}_I$ by a matrix $v : X \times \mathbb{T} \to \mathbb{R}$ as in $(**)$ of Theorem 7. We abuse notation and extend v to specific cases. Formally, we define $v : X \times \mathbb{C} \to \mathbb{R}$ as follows. For $x \in X$ and $c \in \mathbb{C}$, define $v(x, c) = v(x, t)$ for $t \in \mathbb{T} \equiv \mathbb{C}/\sim$ such that $c \in t$. With this definition, $(*)$ of Theorem 1 holds. Obviously, $c \sim d$ implies $v(\cdot, c) = v(\cdot, d)$. The converse also holds: if $v(\cdot, c) = v(\cdot, d)$, $(*)$ implies that $c \sim d$. Finally, observe that, for every distinct four eventualities $x, y, z, w \in X$, the vectors $v(x, \cdot), v(y, \cdot), v(z, \cdot), v(w, \cdot) \in \mathbb{R}^{\mathbb{C}}$ are obtained from the corresponding vectors in $\mathbb{R}^{\mathbb{T}}$ by replication of columns. Since $v : X \times \mathbb{T} \to \mathbb{R}$ is diversified, we also get that $v : X \times \mathbb{C} \to \mathbb{R}$ is diversified.

We now turn to prove that (ii) implies (i). Assume that a diversified matrix $v : X \times \mathbb{C} \to \mathbb{R}$, respecting case equivalence, is given. One may then define $v : X \times \mathbb{T} \to \mathbb{R}$ by $v(x,t) = v(x,c)$ for $t \in \mathbb{T} = \mathbb{C}/\sim$ such that $c \in t$, which is unambiguous because $v(\cdot,c) = v(\cdot,d)$ whenever $c \sim d$. Obviously, $(**)$ of Theorem 7 follows from $(*)$ of Theorem 1, and $v : X \times \mathbb{T} \to \mathbb{R}$ is diversified as well. Defining $\{\succcurlyeq_I\}_I$ by the matrix $v : X \times \mathbb{T} \to \mathbb{R}$ and $(**)$, we find that $\{\succcurlyeq_I\}_I$ satisfy A1*-A4*. Also, $\succcurlyeq_M = \succcurlyeq_{I_M}$ for every $M \in \mathbb{M}$. Hence $\{\succcurlyeq_M\}_M$ satisfy A1-A4.

To see that uniqueness holds, assume that $v, u : X \times \mathbb{C} \to \mathbb{R}$ both satisfy $(*)$ of Theorem 1, and respect case equivalence. Define $v, u : X \times \mathbb{T} \to \mathbb{R}$ as above. The uniqueness result in Theorem 7 yields the desired result. $\square\square$

Step 2: The case of arbitrary $|\mathbb{T}|$ and finite $|X|$.

We first prove that (i) implies (ii). Observe that a representation as in (ii) is guaranteed for every finite $T \subset \mathbb{T}$, provided that T is rich enough to satisfy the diversity axiom A4. We therefore restrict attention to such sets T, and show that the representations obtained for each of them can be "patched" together.

For every ordered list $(x,y,z,w) \in X$, choose $M \in \mathbb{M}$ such that $x \succ_M y \succ_M z \succ_M w$. Such an M exists by A4. Let M_0 be the union of all sets M so obtained. Since X is finite, so is M_0, i.e., $M_0 \in \mathbb{M}$. Let T_0 be the set of types (equivalence classes) of cases in M_0. Choose $g \in X$. Apply Theorem 7 to obtain a representation of $\{\succcurlyeq_I\}_{I \in \mathbb{J}_{T_0}}$ by $v_{T_0} : X \times T_0$ and $(**)$ for all $I \in \mathbb{J}_{T_0} \equiv \mathbb{Z}_+^{T_0}$, such that $v_{T_0}(g,\cdot) = 0$. For every finite $T \subset \mathbb{T}$ such that $T_0 \subset T$, apply Theorem 7 again to obtain a representation of $\{\succcurlyeq_I\}_{I \in \mathbb{J}_T}$ by $v_T : X \times T$ and $(**)$ for all $I \in \mathbb{J}_T \equiv \mathbb{Z}_+^T$, such that $v_T(g,\cdot) = 0$ and such that v_T extends v_{T_0}. v_T is uniquely defined by these conditions. Moreover, if $T \subset T_1 \cap T_2, T_0 \subset T$, and T_1 and T_2 are finite, then the restriction of v_{T_1} and of v_{T_2} to T coincide. The union of $\{v_T\}_{|T|<\infty}$ defines $v : X \times \mathbb{T} \to \mathbb{R}$ satisfying $(**)$ for all $I \in \mathbb{J}_T$ for some finite $T \subset \mathbb{T}$. Defining v on $X \times \mathbb{C}$ as above yields a function that satisfies $(*)$ of Theorem 1 and that respects

case equivalence.

We now turn to prove that (ii) implies (i). Given a representation via a matrix $v : X \times \mathbb{C} \to \mathbb{R}$ as in $(*)$, it follows that $\{\succcurlyeq_M\}_M$ satisfy A1 and A2. A3 also holds since v respects case equivalence. It remains to show that the above, for a diversified v, imply A4. Assume not. Then there are distinct $(x, y, z, w) \in X$ such that for no finite memory M is it the case that $x \succ_M y \succ_M z \succ_M w$. We wish to show that this condition contradicts the fact that v is diversified.

By diversification of v we know that

$$conv\{(v(x,\cdot) - v(y,\cdot)), (v(y,\cdot) - v(z,\cdot)), (v(z,\cdot) - v(w,\cdot))\} \cap \mathbb{R}_-^{\mathbb{C}} = \varnothing.$$

This implies that, for every vector (α, β, γ) in the two-dimensional simplex Δ^2, it is not the case that

$$\alpha(v(x,\cdot) - v(y,\cdot)) + \beta(v(y,\cdot) - v(z,\cdot)) + \gamma(v(z,\cdot) - v(w,\cdot)) \leq 0.$$

In other words, for every $(\alpha, \beta, \gamma) \in \Delta^2$ there exists a case $c \in \mathbb{C}$ such that

$$\alpha(v(x,c) - v(y,c)) + \beta(v(y,c) - v(z,c)) + \gamma(v(z,c) - v(w,c)) > 0.$$

Thus

$$\{\ (\alpha, \beta, \gamma) \in \Delta^2 | \alpha(v(x,c) - v(y,c)) + \beta(v(y,c) - v(z,c)) + \gamma(v(z,c) - v(w,c)) > 0\}_{c \in \mathbb{C}}$$

is an open cover of Δ^2 in the relative topology. But Δ^2 is compact in this topology. Hence it has an open sub-cover. But this implies that there is a finite memory $M \in \mathbb{M}$ such that, restricting v to $X \times M$,

$$conv\{(v(x,\cdot) - v(y,\cdot)), (v(y,\cdot) - v(z,\cdot)), (v(z,\cdot) - v(w,\cdot))\} \cap \mathbb{R}_-^{M} = \varnothing.$$

Let T be the set of types of cases appearing in M. Define $v : X \times T \to \mathbb{R}$ as above. It also follows that

$$conv\{(v(x,\cdot)-v(y,\cdot)),(v(y,\cdot)-v(z,\cdot)),(v(z,\cdot)-v(w,\cdot))\}\cap\mathbb{R}^T_-=\varnothing.$$

By Theorem 7 this implies that there exists $I\in\mathbb{J}_T$ for which $x\succ_I y\succ_I z\succ_I w$. Let M' be a set of cases such that $I(t)=\#(M'\cap t)$, and $M'\subset\cup_{t\in T}t$. It follows that $x\succ_{M'} y\succ_{M'} z\succ_{M'} w$, a contradiction.

Finally, uniqueness follows from the uniqueness result in Step 1. □□

Step 3: The case of infinite X, $\mathbb{T}$.

We first prove that (i) implies (ii). Choose $e,f,g,h\in X$. For $A_0=\{e,f,g,h\}$ there exists a diversified function $v_{A_0}:A_0\times\mathbb{C}\to\mathbb{R}$ satisfying $(*)$ and respecting case equivalence, as well as $v_{A_0}(e,\cdot)=0$. Moreover, all such functions differ only by a multiplicative positive constant. Fix such a function $\widehat{v}_{A_0}$. For every finite set $A\subset X$ such that $A_0\subset A$, there exists a diversified function $v_A:A\times\mathbb{C}\to\mathbb{R}$ satisfying $(*)$ and respecting case equivalence. Moreover, there exists a unique v_A that extends $\widehat{v}_{A_0}$. Let us denote it by $\widehat{v}_A$. We now define $v:X\times\mathbb{C}\to\mathbb{R}$. Given $x\in X$, let A be a finite set such that $A_0\cup\{x\}\subset A$. Define $v(x,\cdot)=\widehat{v}_A(x,\cdot)$. This definition is unambiguous, since, for every two finite sets A_1 and A_2 such that $A_0\cup\{x\}\subset A_1,A_2$, we have $\widehat{v}_{A_1}(x,\cdot)=\widehat{v}_{A_1\cup A_2}(x,\cdot)=\widehat{v}_{A_2}(x,\cdot)$. To see that v satisfies $(*)$, choose $x,y\in X$ and consider $A=A_0\cup\{x,y\}$. Since $v(x,\cdot)=\widehat{v}_A(x,\cdot)$, $v(y,\cdot)=\widehat{v}_A(y,\cdot)$ and $\widehat{v}_A$ satisfies $(*)$ on A, v satisfies $(*)$ on X. Next consider respecting case equivalence, namely, that $v(\cdot,c)=v(\cdot,d)$ iff $c\sim d$. The "if" part follows from the fact that, if $c\sim d$, then for every finite A, $\widehat{v}_A(\cdot,c)=\widehat{v}_A(\cdot,d)$. As for the "only if" part, it follows from the representation by $(*)$ as in Step 1. Finally, to see that v is diversified, let there be given x,y,z,w and choose $A=A_0\cup\{x,y,z,w\}$. Since $\widehat{v}_A$ is diversified, the desired conclusion follows.

The that (ii) implies (i) is follows from the corresponding proof in Step 2, because each of the axioms A1-A4 involves only finitely many eventualities. Finally, uniqueness is proven as in Step 1. □□□

Proof of Proposition 2 – Insufficiency of A1-3:

We show that without the diversity axiom representability is not guaranteed. We provide two counterexamples. The first is combinatorial in nature. The second highlights the role of the diversity axiom in obtaining separability.

Example 1: Let $X = \{ a, b, c, d \}$, $\mathbb{T} = \{1, 2, 3\}$ and $\mathbb{C} = \mathbb{T} \times \mathbb{N}$. Define the vectors in $\mathbb{R}^3$:

$v^{ab} = (-1, 1, 0)$; $\quad v^{ac} = (0, -1, 1)$; $\quad v^{ad} = (1, 0, -1)$;

$v^{bc} = (2, -3, 1)$; $\quad v^{cd} = (1, 2, -3)$; $\quad v^{bd} = (3, -1, -2)$,

and $v^{xy} = -v^{yx}$ and $v^{xx} = 0$ for $x, y \in X$.

For $x, y \in X$ and $I \in \mathbb{Z}^3_+$ define: $x \succsim_I y$ iff $v^{xy} \cdot I \geq 0$. For $M \in \mathbb{M}$, let $I_M \in \mathbb{Z}^3_+$ be the corresponding count vector: $I_M(i) = \#\{(i, j) \,|\, (i, j) \in M\}$ for $i \in \mathbb{T}$, and define $\succsim_M = \succsim_I$.

It is easy to see that with this definition the axioms of continuity and combination, and the completeness part of the order axiom hold. Only transitivity requires a proof. This can be done by direct verification. It suffices to check the four triples (x, y, z) where $x, y, z \in X$ are distinct and in alphabetical order. For example, since $2v^{ab} + v^{bc} = v^{ac}$, $a \succsim_I b$ and $b \succsim_I c$ imply $a \succsim_I c$.

Suppose by way of negation that there are four vectors in $\mathbb{R}^3$, v^a, v^b, v^c, v^d that represent $\succsim_I$for all $I \in \mathbb{J}$ as in Theorem 1. By the uniqueness of representations of half-spaces in $\mathbb{R}^3$, for every pair $x, y \in X$ there is a positive, real number λ^{xy} such that $\lambda^{xy} v^{xy} = (v^x - v^y)$. Further, $\lambda^{xy} = \lambda^{yx}$.

Since $(v^a - v^b) + (v^b - v^c) + (v^c - v^a) = 0$, we have $\lambda^{ab}(-1, 1, 0) + \lambda^{bc}(2, -3, 1) + \lambda^{ca}(0, 1, -1) = 0$. So, $\lambda^{bc} = \lambda^{ca}$, and $\lambda^{ab} = 2\lambda^{bc}$. Similarly, $(v^a - v^b) + (v^b - v^d) + (v^d - v^a) = 0$ implies $\lambda^{ab}(-1, 1, 0) + \lambda^{bd}(3, -1, -2) + \lambda^{da}(-1, 0, 1) = 0$, which in turn implies $\lambda^{ab} = \lambda^{bd}$ and $\lambda^{da} = 2\lambda^{bd}$. Finally, $(v^a - v^c) + (v^c - v^d) + (v^d - v^a) = 0$ implies $\lambda^{ac}(0, -1, 1) + \lambda^{cd}(1, 2, -3) + \lambda^{da}(-1, 0, 1) = 0$. Hence, $\lambda^{ac} = 2\lambda^{cd}$ and $\lambda^{da} = \lambda^{cd}$.

Combining the above equalities we get $\lambda^{ac} = 8\lambda^{ca}$, a contradiction.

Obviously the diversity axiom does not hold. For explicitness, consider the order (b, c, d, a). If for some $I \in \mathbb{J}$, say $I = (k, l, m)$, $b \succ_I c$ and $c \succ_I d$,

then $2k-3l+m > 0$ and $k+2l-3m > 0$. Hence, $4k-6l+2m+3k+6l-9m = 7k - 7\dot{m} > 0$. But $d \succ_I a$ means $m - k > 0$, a contradiction.

The above shows that the relations $\{\succsim_M\}_{M\in\mathbb{M}}$ defined by $\{\succsim_I\}_{I\in\mathbb{Z}^3_+}$ cannot be represented by v that respects equivalence. We now need to show that a matrix v that does not respect case equivalence cannot represent $\{\succsim_M\}_{M\in\mathbb{M}}$ either. Assume that such a matrix existed. Consider memories $M_j = \{(1,j),(2,1)\}$. For every $j \in \mathbb{N}$, $a \approx_{M_j} b$. Hence $v(a,(1,j))$ is independent of j. By similar arguments one shows that v respects case equivalence.

Example 2 : Let $X = [0,1]^2$ and let $\succsim_L$ be the lexicographic order on X. Define, for every non-empty $M \in \mathbb{M}$, $\succsim_M = \succsim_L$, and $\succsim_\varnothing = X \times X$. It is easy to see that $\{\succsim_M\}_{M\in\mathbb{M}}$ satisfy A1-3. However, there cannot be a representation as in $(*)$ since for any non-empty M, $\succsim_M$ is not representable by a real-valued function.□

Proof of Remark 3: In light of the proof of Theorem 1, it suffices to prove the corresponding remark in the framework of Theorem 7.

Assume that $|\mathbb{T}| = 6$. (The proof for $|\mathbb{T}| > 6$ will follow trivially.) Let $\mathbb{T} = \{1,2,3,4,5,6\}$ and $X = \mathbb{R}_+$. We define a matrix $v : \mathbb{T}\times X \to \mathbb{R}$ as follows. For $x \in X$ define the row corresponding to x in the matrix v to be $v_x = (-x^3, -x^2, -x, x, x^2, x^3)$. Define $\succeq_I$ by the matrix v via $(**)$. It suffices to show that v is diversified.

Let there be given distinct $x, y, z, w \geq 0$. Assume that, contrary to diversification, there are nonnegative numbers α, β, γ, with $\alpha + \beta + \gamma = 1$, such that

$$\alpha(v_x - v_y) + \beta(v_y - v_z) + \gamma(v_z - v_w) \leq 0.$$

Comparing the first and the last components of the vector on the left hand side, we obtain

$$\alpha(x^3 - y^3) + \beta(y^3 - z^3) + \gamma(z^3 - w^3) = 0$$

or

$$\alpha x^3 + (\beta - \alpha)y^3 + (\gamma - \beta)z^3 - \gamma w^3 = 0.$$

Similarly, the second and the fifth inequalities yield

$$\alpha x^2 + (\beta - \alpha)y^2 + (\gamma - \beta)z^2 - \gamma w^2 = 0.$$

Finally, the third and fourth inequalities imply

$$\alpha x + (\beta - \alpha)y + (\gamma - \beta)z - \gamma w = 0.$$

Case 1: Assume first that $\alpha \neq 0$. Then there are $\lambda = 1 - \frac{\beta}{\alpha}$, $\mu = \frac{\beta - \gamma}{\alpha}$, and $\nu = \frac{\gamma}{\alpha}$ such that

$$\lambda y^n + \mu z^n + \nu w^n = x^n$$

for $n = 1, 2, 3$. Since we also have $\lambda + \mu + \nu = 1$, it follows that

$$\begin{bmatrix} 1 & 1 & 1 \\ y & z & w \\ y^2 & z^2 & w^2 \\ y^3 & z^3 & w^3 \end{bmatrix} \begin{pmatrix} \lambda \\ \mu \\ \nu \end{pmatrix} = \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \end{pmatrix} \quad .$$
$$(*).$$

By the van der Monde theorem, the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ y & z & w \\ y^2 & z^2 & w^2 \end{bmatrix}$$

is of full rank. Hence, (y^3, z^3, w^3) is spanned by $\{(y^n, z^n, w^n)\}_{n=0,1,2}$. That is, there are $a, b, c \in \mathbb{R}$ such that

$$(y^3, z^3, w^3) = a(1, 1, 1) + b(y, z, w) + c(y^2, z^2, w^2).$$

Combining this equality with $(*)$ yields $x^3 = a + bx + cx^2$. Hence (x^3, y^3, z^3, w^3) is a linear combination of $\{(x^n, y^n, z^n, w^n)\}_{n=0,1,2}$. But this means that the matrix

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ y & z & w & x \\ y^2 & z^2 & w^2 & x^2 \\ y^3 & z^3 & w^3 & x^3 \end{bmatrix}$$

has determinant zero, a contradiction.

Case 2: Assume next that $\alpha = 0$. We then obtain $\beta y^n + (\gamma - \beta) z^n - \gamma w^n = 0$ for $n = 1, 2, 3$. That is,

$$\begin{bmatrix} 1 & 1 & 1 \\ y & z & w \\ y^2 & z^2 & w^2 \\ y^3 & z^3 & w^3 \end{bmatrix} \begin{pmatrix} \beta \\ \gamma - \beta \\ -\gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} .$$

In particular, $(\beta, \gamma - \beta, -\gamma)$ is a solution to the system

$$\begin{bmatrix} 1 & 1 & 1 \\ y & z & w \\ y^2 & z^2 & w^2 \end{bmatrix} \begin{pmatrix} \beta \\ \gamma - \beta \\ -\gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} .$$

Due to the van der Monde theorem again, this can only hold if $(\beta, \gamma - \beta, -\gamma) = (0, 0, 0)$, i.e., if $\beta = \gamma = 0$, in contradiction to the assumption that $\alpha + \beta + \gamma = 1$.□

Proof of Remark 4:

Consider an example in which $\{\succsim_M\}_M$ rank eventualities by relative frequencies, with a tie-breaking rule that is reflected by small additions to the value of v. These small additions, however, vary from case to case and their sum converges. Specifically, let $X = \{1, 2, 3, 4\}$. Define $\mathbb{T} = \{1, 2, 3, 4\}$. $\mathbb{T}$ will indeed end up to be the set of types of cases, as will become clear once we define $\{\succsim_M\}_M$. For the time being we will abuse the term and will refer to elements of $\mathbb{T}$ as "types". Let the set of cases be $\mathbb{C} \equiv \mathbb{T} \times \mathbb{N}$. We now turn to define $v : X \times \mathbb{C} \to \mathbb{R}$. For $x \in X$, $t \in \mathbb{T}$, and $i \in \mathbb{N}$, if $x \neq t$, $v(x, (t, i)) = 0$. Otherwise (i.e., if $x = t$), if $x \in \{1, 2, 3\}$, then $v(x, (t, i)) = 1$. Finally, $v(4, (4, i)) = 1 + \frac{1}{2^i}$ for $i \in \mathbb{N}$. Define $\{\succsim_M\}_M$ by v via $(*)$.

We claim that two cases $(t,i),(s,j) \in \mathbb{T} \times \mathbb{N}$ are equivalent ($(t,i) \sim (s,j)$) iff $t = s$. It is easy to see that if $t \neq s$, then (t,i) and (s,j) are not equivalent. (For instance, $t \succ_{\{(t,i)\}} s$ but $s \succ_{\{(s,j)\}} t$.) Moreover, if $t = s \in \{1,2,3\}$, then $v(\cdot,(t,i)) = v(\cdot,(s,j))$. By $(*)$, $(t,i) \sim (s,j)$. It remains to show that, for all $i,j \in \mathbb{N}$, $(4,i) \sim (4,j)$ despite the fact that $v(\cdot,(4,i)) \neq v(\cdot,(4,j))$.

Observe, first, that $\{\succsim_M\}_M$ agree with relative frequency rankings. Specifically, consider a memory $M \in \mathbb{M}$. Let $I_M \in \mathbb{Z}_+^4$ be defined by $I_M(t) = \#\{\, i \in \mathbb{N} \,|\, (t,i) \in M \,\}$ for $t \in \{1,2,3,4\}$. For any $s,t \in \{1,2,3,4\}$, if $I_M(t) > I_M(s)$, it follows that $t \succ_M s$. Also, if $I_M(t) = I_M(s)$ and $s,t < 4$, then $t \approx_M s$. Finally, if, for $t \in \{1,2,3\}$, $I_M(t) = I_M(4)$, then $4 \succ_M t$.

Let there be given $M \in \mathbb{M}$ such that $(4,i),(4,j) \notin M$. The memories $M \cup \{(4,i)\}$ and $M \cup \{(4,j)\}$ agree on relative frequencies of the types, that is, $I_{M\cup\{(4,i)\}} = I_{M\cup\{(4,j)\}}$. Hence $\succsim_{M\cup\{(4,i)\}} = \succsim_{M\cup\{(4,j)\}}$ and $(4,i) \sim (4,j)$ follows.

Thus v satisfies $(*)$ but does not respect case equivalence.[11] $\square$

Proof of Proposition 5: Assume a representation of $\{\succeq_M\}_{M\in\mathbb{M}}$ by $v = v_y$ as in $(*)$ of Theorem 1 that respects case equivalence. It is easy to see that A5 is necessary for the representation we seek. To prove sufficiency, we first claim that for every $x \in \mathbb{R}^m$, every distinct $a,b,d \in A$, and every $i \in \mathbb{N}$, $v_y(a,(x,d,i)) = v_y(b,(x,d,i))$. To see this, assume, to the contrary, that $v_y(a,(x,d,i)) > v_y(b,(x,d,i))$. Consider M such that $b \succ_M a$ (such an M exists due to the diversity axiom). Adding enough cases of type (x,d) to M will generate a memory M' such that, by $(*)$, $a \succ_{M'} b$, in contradiction to A5.

Since the column vectors $v_y(\cdot,(x,d,i))$ may be shifted by a possibly different constant for each type of cases (x,d), there exists a representation v_y such that $v_y(a,(x,b,i)) = 0$ for $a \neq b$. Moreover, this representation

[11] Observe that the relations $\{\succsim_M\}_M$ satisfy A1 and A2 (as they do whenever they are defined by some v via $(*)$), as well as A4, but not A3. Indeed, such an example cannot be generated if A3 holds as well. Specifically, one can prove the following result: if $\{\succsim_M\}_M$ are defined by v via $(*)$, and satisfy A3 and A4, then $v(x,c) - v(y,c) = v(x,d) - v(y,d)$ whenever $c \sim d$. If, for instance, $v(e,\cdot) \equiv 0$ for some $e \in X$, then v respects case equivalence.

is unique up to multiplication by a positive scalar. It remains to define $s_y(x,a) \equiv v_y(a,(x,a,i))$ for some $i \in \mathbb{N}$. $\square$

Proof of Proposition 6: A6 is obviously implied by the numerical representation we seek. We turn to prove that it is also sufficient. For $x \in \mathbb{R}^m$ consider the memory $M = \{(x,a,1) \,|\, a \in A\}$. The symmetry axiom implies that $a \sim_M b$ for every $a, b \in A$. But this is possible only if $s_y(x,a) = s_y(x,b)$ for every $a, b \in A$. $\square$

References

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

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

de Finetti, B. (1937), "La Prevision: Ses Lois Logiques, Ses Sources Subjectives",*Annales de l'Institute Henri Poincare*, **7**: 1-68.

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

Devroye, L., L. Gyorfi, and G. Lugosi (1996), *A Probabilistic Theory of Pattern Recognition*, New York: Springer-Verlag.

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.

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

Forsyth, R., and R. Rada (1986), *Machine Learning: Applications in Expert Systems and Information Retrieval*, New-York: John Wiley and Sons.

Gilboa, I. and D. Schmeidler (1995), "Case-Based Decision Theory", *Quarterly Journal of Economics*, **110**: 605-639.

Gilboa, I., and D. Schmeidler (1997), "Act Similarity in Case-Based Decision Theory", *Economic Theory*, **9**: 47-61.

Gilboa, I. and D. Schmeidler (1999), "Cognitive Foundations of Probability", Foerder Institute for Economic Research Working Paper No. 30-99.

Gilboa, I. and D. Schmeidler (2001), *A Theory of Case-Based Decisions*, manuscript.

Gilboa, I., D. Schmeidler, and P. P. Wakker (1999), "Utility in Case-Based Decision Theory", Foerder Institute for Economic Research Working Paper No. 31-99.

Hacking, I. (1975), *The Emergence of Probability.* Cambrdige, Cambrdige University Press.

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

Myerson, R. B. (1995), "Axiomatic Derivation of Scoring Rules Without the Ordering Assumption", *Social Choice and Welfare*, **12**, 59-74.

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.

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.

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.

Smith, J. H. (1973), "Aggregation of Preferences With Variable Electorate", *Econometrica*, **41**, 1027-1041.

Stone, C. (1977), "Consistent Nonparametric Regression", *Annals of Statistics* **5**: 689-705.

Young, H. P. (1975), "Social Choice Scoring Functions", *SIAM Journal of Applied Mathematics*, **28**: 824-838.