University of Heidelberg

Department of Economics

Discussion Paper Series | No. 470

Multiple Priors as Similarity Weighted Frequencies

Jürgen Eichberger and
Ani Guerdjikova

July 2008

Multiple Priors as Similarity Weighted Frequencies[1]

Jürgen Eichberger[2] and Ani Guerdjikova[3]

26 May 2008

In this paper, we consider a decision-maker who tries to learn the distribution of outcomes from previously observed cases. For each observed sequence of cases the decision-maker predicts a set of priors expressing his beliefs about the underlying probability distribution. We impose a version of the concatenation axiom introduced in BILLOT, GILBOA, SAMET AND SCHMEIDLER (2005) which insures that the sets of priors can be represented as a weighted sum of the observed frequencies of cases. The weights are the uniquely determined similarities between the observed cases and the case under investigation.

[1] We would like to thank Larry Blume, Alain Chateauneuf, David Easley, Itzhak Gilboa, Joe Halpern, Jean-Yves Jaffray, Marcin Peski and Clemens Puppe, as well as the participants of the Conference on Risk, Uncertainty and Decisions in Tel Aviv 2007, the ESEM in Budapest 2007 and seminar participants at Heidelberg, Cergy-Pontoise, Cornell, Melbourne, and the University of Queensland for helpful comments and suggestions. Financial support from the DFG (SFB 504) and from the Center for Analytic Economics at Cornell is gratefully acknowledged.

[3] University of Heidelberg, Alfred Weber Institute, Grabengasse 14, 69117 Heidelberg, Germany, e-mail: juergen.eichberger@awi.uni-heidelberg.de

[3] Corresponding author. Cornell University, Department of Economics, 462 Uris Hall, e-mail: ag334@cornell.edu.

1 Introduction

How will existing information influence probabilistic beliefs? How do data enter the inductive process of determining a prior probability distribution? KEYNES (1920) discusses in great detail the epistemic foundations of probability theory. In particular, in Part III of his "A Treatise on Probability", he critically reviews most of the then existing inductive arguments for this probability-generating process. One can view the approach of BILLOT, GILBOA, SAMET AND SCHMEIDLER (2005) as an attempt to model this inductive process with the concept of a similarity function, covering both Bayesian and frequentist arguments.

The frequentist approach and the Bayesian belief-based approach to probability theory use available information differently. Both approaches lead, however, to similar statistical results if data are derived from statistical experiments, which are explicitly designed to obtain control over the data-generating process. Classical examples are urn experiments where balls of different colors are drawn from urns with unknown proportions of balls with different colors.

Statistical experiments with identically repeated trials represent an ideal method of data collection. In this case, decision makers can aggregate information directly into a probability distribution over unknown states.

In most real-life decision problems, however, decision makers do not have available data derived from explicitly designed experiments with sufficiently many identical repetitions. Usually, they face the problem to predict the outcome of an action based on a set of data which may be more or less adequate for the decision problem under consideration. This requires aggregating data with different degree of relevance. The case-based decision making approach of GILBOA AND SCHMEIDLER (2001) offers a systematic way to deal with this information aggregation problem: to evaluate an action, the outcomes of past observations are summed up, weighted by their perceived degree of relevance, their *similarity* to the current decision situation.

In a recent paper, BILLOT, GILBOA, SAMET AND SCHMEIDLER (2005), henceforth BGSS (2005), show that, under few assumptions, a probability distribution over outcomes can be derived as a similarity-weighted average of the frequencies of observed cases. Moreover, GILBOA, LIEBERMAN AND SCHMEIDLER (2004) demonstrate how one can estimate the similarity weights from a given database.

The case-based approach in BGSS (2005) associates a database with a single probability distri-

bution. This appears reasonable if the database is large and if the cases recorded in the database are clearly relevant for the decision problem under consideration. Indeed, BGSS (2005) note also that this approach

> "... might be unreasonable when the entire database is very small. Specifically, if there is only one observation, [....] However, for large databases it may be acceptable to assign zero probability to a state that has never been observed." (BGSS (2005), p. 1129)

In order to deal with this problem it appears desirable to choose an approach which allows us to include some notion of ambiguity about the probability distribution associated with a given database. For small and heterogeneous databases ambiguity may be large, while it may vanish for large and homogeneous databases. The multiple-prior approach to decision making offers a framework which captures such ambiguity. Even if a decision maker considers a specific probability distribution as most likely based on the information contained in the data, there may be probability distributions which the decision maker may not want to rule out completely. For example, a decision maker may not trust the information that balls are drawn from an urn with equal numbers of black and white balls. Based on a database consisting of three draws yielding one "black" and two "white" balls, the decision maker may feel ambiguity about whether the probability is 0.5 for the two colors or whether there is a higher probability of a "white" draw. This ambiguity may shrink as the database gets larger and one can be more confident that the proportions of "black" and "white" draws reflect the actual probabilities.

In this paper we modify the approach of BGSS (2005) such that it is possible to consider the weight of increasing evidence. Given a database, we model ambiguity about the most likely probability distribution by a set of probability distributions. We relax the main axiom of BGSS (2005), *Concatenation,* to capture the idea that small data-sets may represent more ambiguous information about the actual probability distribution. At the same time, our modification maintains the main property of the representation derived in BGSS (2005): the similarity function is unique and independent of the content and the size of the data-set.

We then proceed to study more closely the relationship between the amount of data available and the precision of the probabilistic predictions (the size of the set of probability distributions). We assume that the confidence of the decision maker increases as data accumulate and that the set of probability distributions converges to the observed frequency when the data-set becomes large and characterize the similarity function for this situation.

As in BGSS (2005), the question remains open which decision criterion one should use given

the decision maker's beliefs. In order to obtain a decision rule together with a multiple prior representation one may embed these ideas in a behavioral model in the spirit of GILBOA, SCHMEIDLER & WAKKER (2002) or derive decision criteria reflecting degrees of optimism or pessimism in the face of ambiguity as in the work of COIGNARD AND JAFFRAY (1994) and GONZALES AND JAFFRAY (1998). We pursue this approach in EICHBERGER AND GUERDJIKOVA (2008).

There are several ways to model ambiguity of a decision maker in the literature: a representation of ambiguous beliefs by means of capacities was introduced by SCHMEIDLER (1989). BEWLEY (1986)'s approach is based on incomplete preferences. The paper of KLIBANOFF, MARINACCI AND MUKERJI (2005) models ambiguity as second-order risk with respect to the probability distribution determining the outcome. The multiple-prior approach developed by GILBOA AND SCHMEIDLER (1989) generalized by GHIRARDATO, MACCHERONI AND MARINACCI (2004) and CHATEAUNEUF, EICHBERGER AND GRANT (2007) represents ambiguity by a set of probability distributions which a decision maker considers when evaluating her expected utility. In the spirit of these models, we model ambiguity by a set of probability distributions over outcomes. The degree of ambiguity can be measured by set inclusion. The smaller the set of probability distributions over outcomes, the less ambiguous the prediction.

While in the setting of GILBOA AND SCHMEDLER (1989), the set of priors is purely subjective, several recent papers, AHN (2008), GAJDOS, HAYASHI, TALLON AND VERGNAUD (2007), STINCHCOMBE (2003) provide a framework to analyze decisions in situations, in which the set of priors is objectively given, which allows them to separate the objectively given Knightian uncertainty from the subjective attitude towards ambiguity. In our framework, the decision maker associates with each data-set a set of probability distributions, which takes into account the objective information contained in the data (i.e. the nature and frequency of cases observed, as well as the number of observations) and combines it with the subjective attitude of the decision maker towards ambiguity. Hence, our approach provides a method to endogenize the relevant set of priors and connect it to the data-generating process.

EPSTEIN AND SCHNEIDER (2007) analyze statistical learning in the context of ambiguity. Their approach distinguishes between two types of scenarios: those, in which it is possible to learn the objective probability distribution over outcomes and, thus, in the limit the ambiguity disappears, and scenarios, in which the ambiguity persists even in the limit. Our framework

distinguishes between controlled statistical experiments, and situations in which relevant, but not completely identical cases have been observed. We postulate that the decision-maker will be able to learn the objective probability distribution in controlled statistical experiments satisfying the ergodicity property. However, when cases are distinct from the situation under consideration, the decision maker might entertain a set of probability distributions, even in the limit, when a large number of data has been collected.

GONZALES AND JAFFRAY (1998) model preferences over Savage-type acts for a given set of, possibly imprecise, data. They derive a representation of preferences in form of a linear combination of the maximal and the minimal potential outcome of an act and its expected utility with respect to the observed frequency of states. The weights attached to the maximal and minimal outcomes can be interpreted as degrees of optimism and pessimism. They decrease over time relative to the weight attached to the expected utility part of the representation. Because observations may be imprecise a decision maker associates with a set of data a set of priors centered around the observed frequency. The size of the set of probabilities depends negatively on the amount of data.

While we do not derive a decision rule from behavior, our approach encompasses a richer class of situations which allows for, but is not restricted to, the case of controlled statistical experiments considered in both COIGNARD AND JAFFRAY (1994) and GONZALES AND JAFFRAY (1998). The concept of similarity allows us to consider also heterogenous data.

The remainder of the paper is organized as follows. Section 2 presents the model and Section 3 provides some motivating examples. In Section 4 we state the axioms and derive the main representation result, Theorem 4.1. Section 5 deals with the special case of data collected in controlled experiments and Section 6 concludes the paper. All proofs are collected in the Appendix.

2 The Model

The basic element of a *database* is a *case* which consists of an *action* taken and the *outcome* observed together with information about *characteristics* which the decision maker considers as relevant for the outcome. We denote by X a *set of characteristics*, by A a set of *actions,* and by R a set of *outcomes*. All three sets are assumed to be finite. A case $c = (x; a; r)$ is an element of the finite set of cases $C = X \times A \times R$. A *database* of length T is a sequence of

cases indexed by $t = 1...T$:

$$D = ((x_1; a_1; r_1), ..., (x_T; a_T; r_T)) \in C^T.$$

The set of all *databases of length* T is denoted by $\mathbb{D}_T := C^T$.

We also assume that the decision maker is able to make predictions based on the hypothetical evidence of a data-set of infinite length with well-defined frequency. As in the St. Petersburg paradoxon one can consider thought experiments producing potentially infinite data sets. Thus, while it might be impossible to observe such data-sets in practice, the assumption that people are able to make predictions based on such data-sets appears natural. We, therefore, define[4]:

$$\mathbb{D}_\infty = \left\{ (c_t)_{t=1}^{\infty} \mid \text{for each } c \in C, \lim_{T\to\infty} \frac{|\{t \mid c_t = c, \ t \leq T\}|}{T} \text{ exists} \right\}$$

The set of data sets of any length is $\mathbb{D} := \bigcup_{T\in\{1,2...\infty\}} \mathbb{D}_T$.

Consider a decision maker with a given database of previously observed cases, D, who wants to evaluate the uncertain outcome of an action $a_0 \in A$ given relevant information about the environment described by the characteristics $x_0 \in X$. Based on the information in the database D, the decision maker will form a belief about the likelihood of the outcomes. We will assume that the decision maker associates a set of probability distributions over outcomes R,

$$H(D \mid x_0; a_0) \subset \Delta^{|R|-1},$$

with the action a_0 in the situation characterized by x_0 given the database $D \in \mathbb{D}$.

Formally, $H : \mathbb{D} \times X \times A \to \Delta^{|R|-1}$ is a correspondence which maps $\mathbb{D} \times X \times A$ into compact and convex subsets of $\Delta^{|R|-1}$. As usual, the convex combination of two sets of probability distributions H and H' is defined by $\lambda H + (1-\lambda) H' = \{\lambda h + (1-\lambda) h' \mid h \in H \text{ and } h' \in H'\}$. Elements of this set are denoted by $h(D \mid x_0; a_0)$ and we write $h_r(D \mid x_0; a_0)$ for the probability assigned to outcome r by the probability distribution $h(D \mid x_0; a_0)$.

We interpret $H(D \mid x_0; a_0)$ as the set of probability distributions over outcomes which the decision maker takes into consideration given the database D. In contrast to BGSS (2005) we do not assume that a decision maker can always make a point prediction for the probability distribution over outcomes. LUCE AND O'HAGAN (2003) describe the difficulty of making point predictions about probabilities as follows:

> "The first difficulty we will face is that the expert will almost certainly not be an expert in probability and statistics. That means it will not be easy for this person to express her beliefs in the kind of probabilistic form demanded by Bayes' theorem. Our expert may

[4] Lemma A2 in in the Appendix shows how this set can be obtained as a limit of finite data-sets.

be willing to give us an estimate of the parameter, but how do we interpret this? Should we treat it as the mean (or expectation) of the prior distribution, or as the median of the distribution, or its mode, or something else? [...] We could go on to elicit from the expert some more features of her distribution, such as some measure of spread to indicate her general level of uncertainty about the true value of the parameter." (pp. 64-65).

The difficulties of eliciting and interpreting statements about probabilities suggest that, in general, decision makers will not be able to make point predictions about a prior distribution over outcomes. They may, however, identify a range of possible probabilities, either directly as upper and lower bounds of probabilities, or indirectly by a degree of confidence expressed regarding a point prediction. In the former case, a convex set of probabilities is suggested directly, in the latter case, one may view the set of probabilities as a neighborhood of an imprecise point prediction[5].

Like BGSS (2005) we do not explain in this paper how a decision maker chooses an action given the predicted set of priors $H(D \mid x_0; a_0)$. A natural decision criterion would be the minimum expected utility approach introduced by GILBOA AND SCHMEIDLER (1989). We do not suggest a fully behavioral model for a decision criterion and the beliefs in this paper[6]. A characterization of the mapping H from data-sets to probabilities over outcomes is, however, desirable in its own right. It opens up the possibility to study the optimal use of data for the derivation of a set of prior distributions over outcomes for some, not necessarily behavioral, decision criterion.

Notice that these probabilities over outcomes depend both on the action a_0 and the characteristics x_0 of the situation under consideration. In this paper, we will focus on how a decision maker evaluates data in a given decision situation $(x_0; a_0)$. Hence, the *decision situation* $(x_0; a_0)$ will mostly remain fixed. For notational convenience, we will therefore often drop these variables and write simply $H(D)$, $h(D)$ and $h_r(D)$ instead of $H(D \mid x_0; a_0)$, $h(D \mid x_0; a_0)$, and $h_r(D \mid x_0; a_0)$, respectively.

3 Motivating Examples

The following examples illustrate the broad field of applications for this framework. They will

[5] In the Savage framework, GONZALES & JAFFRAY (1998) show how imprecise information may lead to multiple priors. Their paper derives also a representation of the decision maker's preferences in this context.

[6] As pointed out in the introduction, a behavioural foundation of both a decision criterion and a belief correspondence similar to the one proposed in this paper is provided in Eichberger and Guerdjikova (2008).

also highlight the important role of the decision situation $(x_0; a_0)$.
The first example is borrowed from BGSS (2005).

Example 3.1 Medical treatment

A physician must choose a treatment $a_0 \in A$ for a patient. The patient is characterized by a set of characteristics $x_0 \in X$, e.g., blood pressure, temperature, age, medical history, etc. Observing the characteristics x_0 the physician chooses a treatment a_0 based on the assessment of the probability distribution over outcomes $r \in R$. A set of cases D observed [7] in the past may serve the physician in this assessment of probabilities over outcomes.
A case is a combination of a patient t's characteristics x_t, treatment assigned a_t and outcome realization r_t recorded in the database D. Given the database D, the physician considers a set of probabilities over outcomes, $H(D \mid x_0; a_0) \subset \Delta^{|R|-1}$, as possible. These probability distributions represent beliefs about the likelihood of possible outcomes after choosing a treatment a_0 for the patient with characteristics x_0.
Note that we allow the physician to form his beliefs based on cases in which characteristics potentially different from x_0 and actions potentially different from a_0 were observed. E.g., LUCE AND O'HAGAN (2003, PP. 62-64) *discuss how information from different studies about the effectiveness of similar, but not identical, drugs can be combined into a prior distribution. Their example illustrates also why one may want to consider sets of probability distributions, rather than point predictions, as a decision maker's forecast.* ■

A different field of applications are recommender systems which become increasingly popular in internet trade. Internet shops often try to profile their customers in order to provide them with individually tailored recommendations. Our second example shows how an internet provider of a movie rental system can be modelled with this approach.

Example 3.2 Recommender system of an internet movie rental shop

Consider a customer who logs into the internet shop of a movie rental business. The customer is associated with a set of characteristics $x_0 \in X$ which may be more or less detailed depending

[7] The "observations" of cases are not restricted to personal experience. Published reports in scientific journals, personal communications with colleagues and other sources of information may also provide information about cases.

on whether she is a new or a returning customer. The recommender system of the shop has to choose which category of movies a_0 to recommend to this customer. There may be many different categories in an actual recommender system. In this example, we will distinguish, however, only the genre of the movie and the most preferred language of the customer, i.e.,

$$A = \{\textit{Comedy, Documentary, Romance}\} \times \{\textit{English, German}\}.$$

The outcome of the recommendation could be whether rental agreement will result or not, $r \in R = \{$success, no success$\}$.

The recommender system is built on a database D containing records of customers with a profile of characteristics $x_t \in X$ who had been offered a movie $a_t \in A$. Given this database D the system assesses the likelihood $H(D \mid x_0; a_0)$ of the customer x_0 renting a movie from the suggested category a_0. The set of probability distributions over R, $H(D \mid x_0; a_0)$, which are taken into consideration reflects the degree of confidence with respect to this customer. For a new customer, confidence may be low and the set of probabilities $H(D \mid x_0; a_0)$ large. Alternatively, if there are many observations for a returning customer in the database, the set $H(D \mid x_0; a_0)$ may be small, possibly even a singleton. ■

As a final case we will consider a classic statistical experiment where the decision maker bets on the color of the ball drawn out of an urn.

Example 3.3 Lotteries

Consider three urns *with* black *and* white *balls. There may be different information about the composition of these urns. For example, it may be known that*

- *there are* 50 black *and* 50 white balls *in* urn 1,
- *there are* 100 black or white balls *in* urn 2,
- *there is an* unknown number of black and white balls *in* urn 3.

We will encode all such information in the number of the urn, $x \in X = \{1; 2; 3\}$.

In each period a ball is drawn from one of these urns. A decision maker can bet on the color of the ball drawn, $\{B; W\}$. Assume that a decision maker knows the urn x_0 from which the ball is drawn, when he places his bet a_0. An action is, therefore, a choice of lottery *$a \in A := \{1_B 0; 1_W 0\}$, with the obvious notation $1_E 0$ for a lottery which yields $r = 1$ if E occurs and $r = 0$ otherwise.*

Suppose the decision maker learns after each round of the lottery the result and *the urn from which the ball was drawn. Since there are only two possible bets* $a = 1_B 0$ *or* $a' = 1_W 0$ *we can identify cases* $c = (x; a; r)$ *by the urn* x *and the color drawn* B *or* W*. Hence, there are only six cases*

$$C = \{(1; B); (1; W); (2; B); (2; W); (3; B); (3; W)\}.$$

Note that for a given urn x*, the observation of a case, allows the decision maker to observe the outcome of the actually chosen action, but also to infer the (counterfactual) outcome of the lottery he did not choose. This is a specific feature of this example, which distinguishes it from Examples 3.1 and 3.2.*

Suppose that, after T *rounds, the decision-maker has a database*

$$D = ((1; B); (3; W); ...; (2; B)) \in C^T.$$

With each database D*, one can associate a set of probability distributions over the color of the ball drawn* $\{B; W\}$ *or, equivalently, over the payoffs* $\{1; 0\}$ *given a bet* a*. Suppose a decision maker with the information of database* D *has placed the bet* $a_0 = 1_B 0$ *and learns that a ball will be drawn from* urn 2*, then he will evaluate the outcome of this bet based on the set of probability distributions* $H(D \mid 2; a_0)$*. This set should reflect both the decision maker's information contained in* D *and the degree of confidence held in this information. For example, as in statistical experiments, the decision maker could use the relative frequencies of* B *and* W *drawn from* urn 2 *in the database* D *and ignore all other observations in the database. Depending on the number of observations of draws from urn 2, say* $T(2)$*, recorded in the database* D *of length* T*, the decision maker may feel more or less confident about the accuracy of these relative frequencies. Such ambiguity could be expressed by a neighborhood* ε *of the frequencies* $(f_D(2; B); f_D(2; W))$ *of black and white balls drawn from urn* 2 *according to the records in the database* D*. The neighborhood will depend on the number of relevant observations* $T(2)$*, e.g.,*

$$H(D \mid 2; a_0) = \left\{ (h_W; h_B) \in \Delta^1 \mid \ f_D(2; W) - \frac{\varepsilon}{T(2)} \leq h_W \leq f_D(2; W) + \frac{\varepsilon}{T(2)} \right\}.$$

The set of probabilities over outcomes $H(D \mid 2; a_0)$ *may shrink with an increasing number of relevant observations.* ■

The last example illustrates how information in a database may be used and how one can model

ambiguity about the probability distributions over outcomes. In this example, we assumed that the decision maker ignores all observations which do not relate to urn 2 directly. If there is little information about draws from urn 2, however, a decision maker may also want to consider evidence from urn 1 and urn 3, possibly with weights reflecting the fact that these cases are less relevant for a draw from urn 2[8]. The representation derived in the next section allows for this possibility.

4 Axioms and Representation

In this section, we will take the decision situation $(x_0; a_0)$ as given. We will relate the frequencies of cases in a database $D \in \mathbb{D}_T$,

$$f_D(c) := \frac{|\{c_t \in D \mid c_t = c\}|}{T},$$

to sets of probabilities over outcomes $H(D \mid x_0; a_0)$. In particular, let $H_T(D \mid x_0; a_0)$ be the restriction of $H(D \mid x_0; a_0)$ to databases of length T. We will impose axioms on the set of probability distributions over outcomes $H(D \mid x_0; a_0)$ which will imply a representation of the following type: for each $T \in \{2...\infty\}$ and each database of length T,

$$H_T(D \mid x_0; a_0) = \left\{ \frac{\sum_{c \in D} s(c \mid x_0; a_0) f_D(c) \hat{p}_T(c)}{\sum_{c \in D} s(c \mid x_0; a_0) f_D(c)} \mid \hat{p}_T(c) \in \hat{P}_T(c \mid x_0; a_0) \right\}.$$

The set of probability distributions over outcomes $\hat{P}_T(c \mid x_0; a_0)$ denotes the beliefs of the decision maker when the database $D = \left(\underbrace{c...c}_{T\text{-times}}\right)$ is observed. As will be discussed in more detail in Section 5, this set may depend on the number of observations. It may be large for small numbers and may shrink as more confirming data become available.

The weighting function $s(c \mid x_0; a_0)$ represents the relevance of a case c for the current situation $(x_0; a_0)$ and can be interpreted as the perceived similarity between c and $(x_0; a_0)$. We will characterize a similarity function which depends only on the cases compared and not on the number of observations.

Independence from the number of observations is justified if one assumes that the similarity of cases is determined by non-quantifiable knowledge about the cases[9]. Moreover, it is not clear how more observations of a case should affect the similarity between this case and the other cases. If the number of observations increases without affecting the relative frequencies of

[8] Part III of KEYNES (1921) provides an extensive review of the literature on induction from cases to probabilities.

[9] Compare also the discussion of "structural priors" in LUCE & O'HAGAN (2003, PP. 67-68).

cases, then there is no reason to adjust the similarity relation between cases. Notice however that the relative weight given to the outcome probabilities based on the observations of a particular case increases with the relative frequency of this case.

The axioms suggested below will imply unique (up to a normalization) similarity weights $s(c \mid x_0; a_0)$, which do not depend on T and, for each $T \in \{2...\infty\}$, unique sets of probability distributions $\hat{P}_T(c \mid x_0; a_0)$. This result generalizes the main theorem of BGSS (2005) to the case of multiple priors.

In the following discussion, $(x_0; a_0)$ is assumed constant and we suppress notational reference to it. It is important to keep in mind, however, that all statements of axioms and conclusions do depend on the relevant reference situation $(x_0; a_0)$. In particular, the similarity weights, deduced below, measure similarity of cases relative to this reference situation.

In order to characterize the mapping $H(D)$ we will impose axioms which specify how beliefs over outcomes change in response to additional information. In general, it is possible that the order in which data become available conveys important information. We will abstract here from this possibility and assume that only data matter for the probability distributions over outcomes.

Axiom A1 (*Invariance*) Let π be a one-to-one mapping $\pi : \{1...T\} \to \{1...T\}$, then

$$H\left((c_t)_{t=1}^T\right) = H\left(\left(c_{\pi(t)}\right)_{t=1}^T\right).$$

According to Axiom $(A1)$, the set of probability distributions over outcomes is invariant with respect to the sequence in which data arrive. Hence, each database D is uniquely characterized by the tuple $(f_D; |D|)$, where $f_D \in \Delta^{|C|-1}$ denotes the vector of frequencies of the cases $c \in C$ in the data-set D and $|D|$ the length of the database.

Remark 4.1 *By Axiom $(A1)$, we can identify every data-set $D = (c_t)_{t=1}^T$ with the corresponding multi-set $\left\{(c_t)_{t=1}^T\right\}$, in which the number of appearances of every case c exactly corresponds to the number of its appearances in D. We will denote the data-set and its corresponding multi-set by the same letter. In particular, when we write $D = D'$, we mean equality of the multi-sets corresponding to the data-sets D and D'.*

In line with BGSS (2005), we call the combination of two databases a *concatenation*.

Definition 4.1 *(*Concatenation*) For any T, $T' \in N \cup \{\infty\}$, and any two databases $D =$*

$(c_t)_{t=1}^{T}$ *and* $D' = (c'_t)_{t=1}^{T'}$*, the database*

$$D \circ D' = \left((c_t)_{t=1}^{T} ; (c'_t)_{t=1}^{T'} \right)$$

is called the concatenation of D *and* D'*.*

By Axiom $(A1)$, concatenation is a commutative operation on databases. The following notational conventions are useful.

Notation $D^k = \underbrace{D \circ ... \circ D}_{k\text{-times}}$ denotes k concatenations of the same database D. In particular, a database consisting of k-times the same case c can be written as $(c)^k$. We will use D^∞ to denote the infinite data-set with frequency of observations f_D.

Imposing the following *Concatenation Axiom*, BGSS (2005) obtain a characterization of a function h mapping $\mathbb{D}$ into a single probability distribution over outcomes.

Axiom (BGSS 2005) (*Concatenation*) For every D, $D' \in \mathbb{D}$**,**

$$h(D \circ D') = \lambda h(D) + (1 - \lambda) h(D')$$

for some $\lambda \in (0; 1)$.

This axiom can be easily adapted to our framework:

Axiom (BGSS – multiple priors) (*Concatenation with multiple priors*) For every D, $D' \in D$

$$H(D \circ D') = \lambda H(D) + (1 - \lambda) H(D')$$

for some $\lambda \in (0; 1)$.

Both versions of the axiom imply that, for any k, the databases D and D^k map into the same set of probability distributions over outcomes, $H(D) = H(D^k)$. Hence, two data-sets $D = (c)$ and $D' = (c)^{10000}$ will be regarded as equivalent. This seems counterintuitive. Ten thousand observations of the same case $c = (x; a; r)$ are likely to provide stronger evidence for the outcome r in situation $(x; a)$ than a single observation. Arguably, the database $(c)^{10000}$ provides strong evidence for a probability distribution concentrated on the outcome r; $h_r((c)^{10000}) = 1$. Based on a single observation $(x; a; r)$, however, it appears quite reasonable to consider a set of probability distributions $H((c))$ which also contains probability distributions $h((c))$ with $h_{r'}((c)) \in (0,1)$ for all r'. In particular, based on the information contained in $D = (c)$, a decision maker may not be willing to exclude the case of all outcomes being equally probable,

i.e., $\overline{h}(D)$ with $\overline{h}_{r'}(D) = \frac{1}{|R|}$ for all $r' \in R$. It appears perfectly reasonable to include $\overline{h}$ in $H((c))$ but not in $H\left((c)^{10000}\right)$.

Since we would like to capture the fact that confidence might increase as the number of observations grows, we cannot simply apply the *Concatenation Axiom* of BGSS (2005) to all probability distributions in the mapping H. Restricting the axiom to databases with equal length will provide sufficient flexibility for our purpose.

To illustrate the idea, consider two cases c_1 and c_2 and data-sets with two observations of these cases, say $D_1 = (c_1, c_1)$, $D_2 = (c_2, c_2)$, and $F = (c_1, c_2)$. Due to Axiom $(A1)$, one can write these data-sets in terms of frequencies and numbers of observations as $F = (f_F, 2)$, $D_1 = (f_{D_1}, 2)$, and $D_2 = (f_{D_2}, 2)$. Since $D_1 \circ D_2 = (c_1, c_1, c_2, c_2) = F \circ F$ holds, the frequency of cases in F must be a mixture of the frequencies of D_1 and D_2,

$$f_F = \frac{1}{2} f_{D_1} + \frac{1}{2} f_{D_2}.$$

Whatever the predictions $H(D_1)$ and $H(D_2)$, which the decision maker makes based on the databases D_1 and D_2, the prediction for the data-set $F = (c_1, c_2)$ should in some sense lie between $H(D_1)$ and $H(D_2)$. Formally, we will require the existence of a $\lambda \in (0, 1)$ such that $\lambda H(D_1) + (1 - \lambda) H(D_2) = H(F)$.

Restricting the concatenation axiom of BGSS (2005) to data-sets with the same number of cases suffices for predictions to depend both on frequencies and the number of observations. It is not sufficient, however, to guarantee that the similarity function is independent of the number of observations. In order to obtain a similarity function independent of T we need, in addition, that the convex combination be independent of the length T, i.e., $\lambda H\left(D_1^k\right) + (1 - \lambda) H\left(D_2^k\right) = H\left(F^k\right)$ for all $k \geq 1$.

Axiom $(A2)$ generalizes this idea: for any n data-sets of equal length T that can be concatenated to an n-fold of a data-set F of length T, we postulate that any probability distribution over outcomes predicted on the basis of data-set F can be expressed as a convex combination of probability distributions over outcomes associated with the data sets D_i. In addition, the weights should remain invariant to scaling up the data-sets.

Axiom A2 (*Concatenation*) Consider data-sets $F \in \mathbb{D}_T$ and $D_1 ... D_n \in \mathbb{D}_T$ for some $n \in \mathbb{Z}_+$, such that $D_1 \circ ... \circ D_n = F^n$. Then, there exists a vector $\lambda \in int\left(\Delta^{n-1}\right)$ such that, for every

$k \in \mathbb{Z}^+$,

$$\sum_{i=1}^{n} \lambda_i H\left(D_i^k\right) = H\left(F^k\right).$$

In spirit, Axiom $(A2)$ is very similar to the Concatenation Axiom introduced by BGSS (2005). It has the following intuitive interpretation[10]: if a decision maker cannot exclude a certain probability distribution h after observing the evidence in any of the data-sets $D_1 ... D_n$, then he should not be able to exclude it after observing the evidence in F. The main difference to the Concatenation Axiom of BGSS (2005) is that we restrict the axiom to data-sets of equal length.

The restriction to sets of equal length is important for our approach since databases with identical frequencies, but different length may give rise to different sets of probabilities over outcomes. In particular, depending on some learning rule[11], it may be reasonable to assume that the set of probabilities over outcomes shrinks as more observations of the same cases become available. Intuitively, the decision maker is more confident that the observed frequencies reflect the actual data-generating process for the data-set D^{T+1} than for D^T, hence, $H(D^{T+1}) \subset H(D^T)$.In contrast, applying the *Concatenation Axiom* of BGSS (2005), we would have to conclude that for some $\lambda \in int\left(\Delta^T\right)$,

$$\begin{aligned} H\left(D^{T+1}\right) &= H\left(D^T \circ D\right) = \lambda_1 H(D^T) + (1-\lambda_1)H(D) = \\ &= \lambda_1 H(D^{T-1}) + \lambda_2 H\left(D\right) + (1-\lambda_1-\lambda_2)H(D) \\ &= \sum_{i=1}^{T} \lambda_i H\left(D\right) = H\left(D\right). \end{aligned}$$

for all T. Thus, imposing BGSS (2005)'s *Concatenation Axiom*, the set of probability distributions over outcomes would necessarily be independent of the number of observations. Our weaker Axiom $(A2)$, however, implies in this case only $\sum_{i=1}^{T} \lambda_i H\left(D\right) = H\left(D\right)$, which is trivially satisfied for any set D.

Remark 4.2 *Our Axiom $(A2)$ suggests that sets of equal length have identical degrees of confidence. This may be a natural assumption in some applications, but it needs not hold in general. For example, the degree of confidence of a given data-set may depend on its length and on its*

[10] Note that the Axiom *does not have* the following behavioral implication: if action a is preferred to a' under all data-sets $D_1 ... D_n$, then it is also preferred under F. To understand this, consider the case of $n = 2$. Let $a \succsim_{D_1} a'$ and $a \succsim_{D_2} a'$. Suppose also that the evidence contained in the data-set D_1 is more relevant for a, while the evidence contained in D_2 is more relevant for a'. Suppose that, at the same time, the decision-maker values a' higher given the relevant evidence contained in D_2 than he values a, given the relevant evidence for this alternative, D_1. In this case, combining the evidence contained in the two data-sets D_1 and D_2 into F might lead to a reversal of preferences, i.e., $a' \succsim_F a$. The same argument applies also for the Concatenation Axiom of BGSS (2005).

[11] Section 5 discusses learning in more detail.

frequency. For such a situation, one can derive the sets of databases, for which the decision-maker is equally confident in the following way.
Suppose that for a given $T \in \mathbb{Z}_+$, all sets of the type $(c_i)^T$ for $i \in \{1...|C|\}$ provide the same degree of confidence. This seems intuitive, since these databases consist of a T-fold repetition of the same case. While predictions based on this case may differ, the degree of confidence in these predictions should remain constant across the different cases. Consider the following modified concatenation axiom:

Axiom $A2'$ *Let $D \in \mathbb{D}_T$. Then there exists a number $S(D) \in \mathbb{Z}_+$ and a vector $\lambda \in \operatorname{int}\left(\Delta^{|C|-1}\right)$ such that*

$$H(D) = \sum_{i=1}^{|C|} \lambda_i f_D(c_i) H\left(c_i^{S(D)}\right)$$

Furthermore, for each $k \in \mathbb{Z}_+$,

$$H\left(D^k\right) = \sum_{i=1}^{|C|} \lambda_i f_D(c_i) H\left(c_i^{k \cdot S(D)}\right).$$

One may call $S(D)$ the degree of confidence of the data-set D. We can define the sets $\mathbb{D}_S$, i.e., the set of databases with confidence S. Note that the case in which $S(D) = T$ for all $D \in \mathbb{D}_T$ is a special case in which Axiom $A2'$ holds. However, $A2'$ also allows for more general functions $S(\cdot)$. E.g., for some positive numbers A and B, $S(\cdot)$ could be given by:

$$S(D) = A \left\| \left(\underbrace{\frac{1}{|C|} ... \frac{1}{|C|}}_{|C|\text{-times}} \right) - f_D \right\| + B|D|.$$

It is easy to check that this function satisfies the property of Axiom $A2'$. The statement of Theorem 4.1 remains unchanged if we replace Axiom $(\mathrm{A2})$ with Axiom $(\mathrm{A2}')$ and the sets $\mathbb{D}_T$ with $\mathbb{D}_S$.

Similar to BGSS (2005), we have to impose a linear-independence condition on the sets of probability distributions over outcomes $H(D)$.

Axiom A3 (*Linear Independence*) For every $T \in \{2, 3...\infty\}$, the basis of $\mathbb{D}_T$, $(c_1)^T$,...,$\left(c_{|C|}\right)^T$ satisfies the following condition:

There are at least three distinct i, j, $k \in \{1...|C|\}$, such that $H\left((c_i)^T\right)$, $H\left((c_j)^T\right)$ and $H\left((c_k)^T\right)$ are:

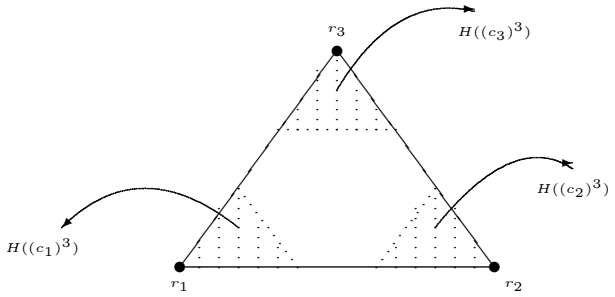
– either singletons

$$H\left((c_m)^T\right) = \left\{h\left((c_m)^T\right)\right\} \text{ for } m \in \{i; j; k\}$$

and $h\left((c_i)^T\right)$, $h\left((c_j)^T\right)$ and $h\left((c_k)^T\right)$ are non-collinear,

– or polyhedra with a non-empty interior such that no three of their extreme points are collinear.

As an example of sets $H(D)$ satisfying Axiom $(A3)$ consider the case of $|C| = |R| = 3$. In particular, take $c_1 = (x; a; r_1)$, $c_2 = (x; a; r_2)$ and $c_3 = (x; a; r_3)$. Suppose that each of the $H\left((c_i)^T\right)$ represents a confidence interval around the actually realized frequency of outcomes, $e_1 = (1; 0; 0)$, $e_2 = (0; 1; 0)$ and $e_3 = (0; 0; 1)$. Then, these sets will satisfy the requirement of Axiom $(A3)$, see Figure 1.



1. Non-collinear sets of priors

The following theorem guarantees a unique similarity function for data-sets of arbitrary length.

Theorem 4.1 *Let H be a correspondence $H : \mathbb{D} \to \Delta^{|R|-1}$ the images of which are non-empty convex and compact sets and let $H_T(D)$ be the restriction of H to $\mathbb{D}_T$. Then the following two statements are equivalent:*

(i) *H satisfies the Axioms* Invariance $(A1)$*,* Concatenation $(A2)$*, and* Linear Independence $(A3)$.

(ii) *There exists a function*

$$s : C \to \mathbb{R}_{++}$$

and a correspondence

$$\hat{P} : \{2, 3...\infty\} \times C \to \Delta^{|R|-1}$$

satisfying Linear Independence *and such that for any $D \in \mathbb{D}_T$ with $T \in \{2, 3...\infty\}$,*

$$H_T(D) = \left\{ \frac{\sum_{c \in C} s(c) \hat{p}_T(c) f_D(c)}{\sum_{c \in C} s(c) f_D(c)} \mid \hat{p}_T(c) \in \hat{P}_T(c) \right\}.$$

Moreover, $\hat{P}$ is unique and s is unique up to a multiplication by a positive number.

Note how the different axioms enter this representation. $(A1)$ insures that the only relevant

characteristics of a data-set D are the generated frequencies $(f_D(c))_{c\in C}$ and its length T. We then use $(A2)$ and $(A3)$ to show that for a class of databases with infinite length, we can represent $H(D)$ as a union of functions $h(D)$ which satisfy the axioms of BGSS (2005). This class of data-sets can be characterized by its frequencies, which are dense in the simplex of dimension $|C|-1$. Hence, we can apply Proposition 3 of BGSS (2005) to every selection $h(D)$ in order to demonstrate the existence of a unique (up to a multiplication by a positive constant) similarity function s and unique probabilities $\hat{p}$. Axiom $(A2)$ then implies that the same values of s can be used for every $T<\infty$.

Note that replacing $(A2)$ by the Axiom (BGSS – multiple priors) in Theorem 4.1 would imply that both, the predictions $H(D)$ and the correspondence $\hat{P}$ depends only on the case c, but not on the length of the data-set, T. In other words, the predictions made by the decision maker depend only on the frequency of cases in a data-set, but not on its length.

5 Learning and confidence

The representation for the prediction correspondence $H(D)$ derived in Theorem 4.1 determines neither the ambiguity captured by the basic prediction sets $\hat{P}_T(c)$ nor the similarity function $s(c)$. These degrees of freedom may be be filled either by a behavioral approach which derives a decision rule and a prediction correspondence from preferences or by more normative considerations about how predictions should be made as statistics.

In this section we will make a first attempt to narrow down the indeterminateness of the basic predictions $\hat{P}_T(c)$ and the similarity function $s(c)$. These restrictions are derived from the assumption that the prediction correspondence $H(D)$ should satisfy some natural conditions when D contains data obtained from controlled statistical experiments. Though these results are not difficult to derive, they do point the way to a more tightly specified learning rule.

Additional information in the form of more data may affect

- the ambiguity about a prediction as represented by the set of probabilities over outcomes,
- the prediction as represented by the type of probabilities considered, and
- the similarity between different cases.

A natural idea about learning is provided by controlled statistical experiments. Any sensible prediction correspondence $H(D)$ should satisfy these conditions for the special case of

data-sets obtained from such experiments. A decision maker who would like to learn the outcome distribution for a given action a and for given characteristics x would like to run a series of experiments creating cases with constant (x, a) in order to learn about the frequencies of outcomes r_t. In statistical experiments, e.g., in the urn experiments of Example 3.3, data-sets $D = ((x, a, r_t))_{t=1}^{T}$ are generated.

Learning a probability distribution is meaningful only if we assume stationarity and ergodicity of the underlying random process according to which the outcome is generated. The learning process of the decision maker begins with a set of probability distributions over outcomes. In the case of a repeated experiment, where $(x; a)$ is constant, the set of probability distributions over outcomes is assumed to contain the actually observed frequencies. Given the assumption of ergodicity, as the data-set becomes larger, the ambiguity of the decision maker decreases until, as the number of observations increases, the set of probability distributions reduces to a singleton. Moreover, if the assumption of ergodicity is satisfied and $D = ((x; a; r_t))_{t=1}^{\infty}$, then, according to the Ergodic Theorem, DURETT (2005, P. 337), the frequencies of r a.s. converge to a distribution $f(r)$ which exactly corresponds to the actual probability distribution of r given $(x; a)$:

$$\lim_{T\to\infty} \frac{|\{t \leq T \mid r_t = r\}|}{T} = \lim_{T\to\infty} f_T(r) = f(r).$$

These considerations motivate the following axiom which characterizes how decision makers may learn a probability distribution over outcomes and gain confidence about their predictions. The axiom seems natural in the context of controlled experiments. It requires that ambiguity decreases as "more and more cases with the same outcome" are observed. Moreover, if the same outcome is observed over and over again, its perceived probability converges to 1.

Axiom A4 *(Learnability)* Consider databases with fixed $(x; a)$, $D = \left\{(x; a; r_t)_{t=1}^{T}\right\}$. As $T \to \infty$,

$$H(D \mid x; a) \to \{h(D \mid x; a)\}$$

with

$$h_r(D \mid x; a) = f_D(r).$$

Axiom $(A4)$ implies that the decision maker can learn the unknown proportion of the colors in a given urn, as in Example 3.3, if the draws from the urn are with replacement, and the number of observations becomes large.

According to Axiom $(A4)$ ambiguity will disappear in the limit. In the context of controlled experiments it appears also reasonable to postulate that ambiguity decreases with the number of observations.

Axiom A5 *(Accumulation of knowledge)* For, $T' > T$, let $D \in \mathbb{D}_T$ and $D' \in \mathbb{D}_T$ be two finite data-sets with identical frequencies $f_D = f_{D'}$. Then

$$H\left(D' \mid x; a\right) \subset H(D \mid x; a).$$

Axiom $(A5)$ captures the idea that the ambiguity of the decision maker about the true probability distribution of r decreases as the number of observations increases. It does not tell us, however, in which way the set of probabilities over outcomes shrinks.

Notice that Axiom $(A5)$ applies only to data-sets with *identical frequencies*. If frequencies differ, a smaller set might provide more reliable information than a larger one. For example, $D \in \mathbb{D}_{100}$ with $f_D\left(x; a; r_1\right) = \frac{99}{100}$ and $f_D\left(x; a; r_2\right) = \frac{1}{100}$ will in general constitute stronger support for $h\left(r_1 \mid x; a\right) = \frac{99}{100}$ than $D' \in \mathbb{D}_{200}$ with $f_{D'}\left(x; a; r_1\right) = f_{D'}\left(x; a; r_2\right) = \frac{1}{2}$.

Together with the *Invariance Axiom* $((A1))$, Axioms $(A4)$ and $(A5)$ imply that the observed frequency of outcomes in a controlled experiment is always contained in the set of probabilities over outcomes which the decision maker considers.

Lemma 5.1 *Assume Axioms* $(A1)$, $(A4)$ *and* $(A5)$ *hold, then for any database* D *of length* T *with fixed* $(x; a)$, *i.e.,* $D = \left((x; a; r_t)_{t=1}^{T}\right)$, *there is an* $h \in H\left(D \mid x; a\right)$ *such that, for all* $r \in R$,

$$h_r\left(D \mid x; a\right) = f_D\left(r\right).$$

Finally, we prove that for the representation derived in Theorem 4.1, Axioms $(A4)$ and $(A5)$ imply two intuitive properties of the representation of $H(D)$. Firstly, the sets $\hat{P}_T\left(x; a; r\right)$ shrink with time, always contain the r-th unit vector e_r and, as T converges to infinity, converge to e_r . Secondly, for a given tuple $(x; a)$, the similarity function assigns a value of 1 (up to a normalization) to all cases $(x; a; r')$ with $r' \in R$. Hence, as long as the conditions under which the experiment is conducted remain constant, all outcomes of the experiment are equally relevant for the assessment of probabilities.

Theorem 5.2 *Suppose Axioms* $(A4)$ *and* $(A5)$ *hold. Then the representation* $H(D \mid x; a)$ *in Theorem 4.1 satisfies the additional properties:*

1. *For all* $r \in R$ *and all* T*,*

   (i) $\hat{P}_T((x;a;r) \mid x;a) \subset \hat{P}_{T-1}((x;a;r) \mid x;a)$*,*

   (ii)$e_r \in \hat{P}_T((x;a;r) \mid x;a)$*, and*

   (iii) $\lim_{T\to\infty} \hat{P}_T((x;a;r) \mid x;a) = \{e_r\}$*,*

   where e_r *denotes the* r*-th unit vector of dimension* $|R|$*.*

2. *For all* $r \in R$*,* $s((x;a;r);(x;a)) = 1$*.*

Statement 1 of Theorem 5.2 follows immediately from the Axioms $(A4)$ and $(A5)$. It is less obvious that these assumptions about learning in a controlled environment do also constrain the similarity function. As Statement 2 shows, the notion of a statistical experiment implies that the similarity function must be independent of the outcomes.

There are however no other restrictions for the similarity function implied. These degrees of freedom need to be specified from knowledge unrelated to the generated data. Such constraints may be imposed by a relation on data-sets which describes their intrinsic similarity according to some criteria or by a behavioral approach which derives the similarity function implied by the preferences of the decision maker[12].

Our model captures different features of learning. First, it allows us to distinguish between two types of ambiguity: (i) ambiguity resulting from the fact that the data-set is not sufficiently long, i.e., uncertainty about whether the realized frequency represents the outcome generating process and, (ii) ambiguity resulting from the fact that not all cases are equally relevant for the prediction to be made. The result of Theorem 5.2 concerns the first type of ambiguity. It states that this ambiguity diminishes and eventually disappears as the number of observations grows. This result, however applies only to controlled statistical experiments. In contrast, if observations of action a' induce predictions about a, then the set of probability distributions

[12] In applications, the former approach is used, e.g., in GUERDJIKOVA (2007), who relates the curvature of the similarity function to preferences for diversification. GUERDJIKOVA (2008) identifies properties of the similarity function which ensure that a case-based decision maker learns to choose the optimal action in the limit. In contrast, GAYER (2007) models a situation, in which similarity considerations are relevant only when the number of observations is small. Hence, she uses a similarity function which converges to the identity function as the data-set becomes large. The behavioral approach is pursued by GILBOA, LIEBERMAN & SCHMEIDLER (2004) who estimate a similarity function from data about the Tel Aviv rental market for apartments.

$\hat{P}_T\left((x;a';r) \mid (x;a)\right)$ need not shrink to a singleton, even for large numbers of observations. This second type of ambiguity may be persistent. If, for instance, the decision maker is uncertain about correlations in the outcomes of actions a and a', even long data-sets may not convey additional information about the correlation structure.

Our model can thus be interpreted as one of learning under model uncertainty. In contrast to the Bayesian model, which has to take all eventualities into account ex ante, our model allows for "surprises". For example, after observing a certain case $(x;a;\bar{r})$ for 1000 times, a decision maker may entertain the set of priors

$$H\left((x;a;\bar{r})^{1000}\right) = \left(h\left(\bar{r}\right) \in [1-\epsilon;1]\,; h\left(r'\right) \in [0;\epsilon]\,; h\left(r''\right) = 1 - h\left(\bar{r}\right) - h\left(r'\right)\right)$$

for some outcomes r' and r'' distinct from $\bar{r}$, and assign probability 0 to all other outcomes. Now assume that the next case observed is $(x;a;\hat{r})$ ($\hat{r} \notin \{r';r'';\bar{r}\}$). From the point of view of the decision maker the occurrence of $\hat{r}$ is a surprise. Our Concatenation Axiom requires the decision maker to put a positive weight on the evidence from this case and, since by Theorem 5.2, $e_{\hat{r}} \in H\left((x;a;\hat{r})^T\right)$ for all T, it follows that after this additional observation, the decision maker will put a positive probability on $\hat{r}$. Hence, the learning process in this model can have properties which differ substantially from standard learning models. Furthermore, if we allow the degree of confidence of D to depend not only on the number but also on the frequency of observations, as in Axiom $2A'$ (Remark 4.2), then the degree of confidence of the new data-set $(x;a;\bar{r})^{1000} \circ (x;a;\hat{r})$ may be smaller that that of $(x;a;\bar{r})^{1000}$. Thus, not only the beliefs of the decision maker may change with surprises, but also the perceived ambiguity as measured by the size of the set $H\left(D\right)$.

6 Concluding remarks

The amount of data available may influence a decision maker's confidence in a probability distribution. In this paper, we combine this intuition with the similarity-weighted frequency approach of BGSS (2005). We relax the *Concatenation Axiom* of BGSS (2005) by restricting it to databases of equal length. We show that the main result of BGSS (2005), namely that the similarity function is unique, remains true if one imposes consistency on the weights across data-sets of different size. This consistency is essential for the uniqueness of the similarity weights.

As a special case of our approach we consider predictions associated with homogenous data-sets

which contain the same characteristics and actions in all observations. Homogenous data-sets can arise from controlled statistical experiments. In this context, it appears natural that ambiguity decreases as new data confirm past evidence. Combined with the assumption that, in the limit, the decision maker learns the probability distribution generating the process, one obtains the conclusion that all observations are considered equally important. Statistical experiments can serve as an illustration of our approach. Similarity becomes important however, when data-sets contain only few cases.

In this paper, we derive a representation in which the similarity weights are independent of the amount of data. If one views the perception of similarity, however, as an imperfect substitute for knowledge about the relevance of underlying data, then a decision maker has to *find out* which characteristics are payoff-relevant. Hence, the data-set may provide not only information about the distribution of payoffs, but also about similarity of alternatives. The more observations a data-set contains, the more precise the perception of similarity may become.

This observation raises several questions for further research. On the one hand, one may try to model the adjustment of the similarity function in the light of new information. PESKI (2007) suggests a possible approach. He describes a learning process, in which the decision-maker tries to assign objects optimally to categories in order to make correct predictions. Two objects do either belong to a category or do not belong to it. One can interpret this approach as a restriction of the similarity values to zero or one. A less restrictive model would allow for a continuum of similarity values.

A second research agenda concerns the derivation of a decision rule and a multiple-prior representation of beliefs from preferences over actions and data-sets. Combining axioms from case-based decision making and from the literature on decision making under ambiguity, it is possible to find a representation of preferences over acts and a set of probabilities over outcomes depending on the data-set. We pursue this issue in EICHBERGER AND GUERDJIKOVA (2008).

Appendix A. Proofs

Proof of Theorem 4.1 :

We first show necessity: it is obvious that for a given $D \in \mathbb{D}_T$,

$$H_T(D) = \left\{ \frac{\sum_{c \in C} s(c)\, \hat{p}_T(c)\, f_D(c)}{\sum_{c \in C} s(c)\, f_D(c)} \mid \hat{p}_T(c) \in \hat{P}_T(c) \right\}$$

does not depend on the order of cases observed in D, but only on their frequency and the length of D, T, hence Axiom $(A1)$ is satisfied. To see that Axiom $(A2)$ is satisfied, first note that for all $c \in C$ and all $T \in \{2, 3...\infty\}$,

$$H_T\left(c^T\right) = \left\{ \frac{s\left(c\right)\hat{p}_T\left(c\right)}{s\left(c\right)} \mid \hat{p}_T\left(c\right) \in \hat{P}_T\left(c\right) \right\} = \hat{P}_T\left(c\right).$$

Choose a data-set $F = (c_1...c_T) \in D_T$ and decompose it as in Axiom $(A2)$ into T sets: $D_1 = c_1^T ... D_T = c_T^T \in D_T$ such that:

$$D_1 \circ ... \circ D_T = F^T.$$

It is obvious that:

$$f_F\left(c_t\right) = \frac{1}{T}\sum_{i=1}^{n} f_{D_i}\left(c\right)$$

and we have:

$$\begin{aligned} \frac{\sum_{c\in C} s\left(c\right)\hat{p}_T\left(c\right) f_F\left(c\right)}{\sum_{c\in C} s\left(c\right) f_F\left(c\right)} &= \sum_{c\in C} \frac{s\left(c\right) f_F\left(c\right)}{\sum_{c\in C} s\left(c\right) f_F\left(c\right)} \hat{p}_T\left(c\right) = \\ \sum_{t=1}^{T} \frac{s\left(c_t\right)}{\sum_{t=1}^{T} s\left(c_t\right)} \hat{p}_T\left(c\right) &= \sum_{t=1}^{T} \lambda_t \cdot \hat{p}_T\left(c\right) \end{aligned}$$

with $\lambda_t = \frac{s(c_t)}{\sum_{t=1}^{T} s(c_t)}$. Note that since $s\left(c_t\right) > 0$ for all $c_t \in C$, $\left(\lambda_t\right)_{t=1}^{T} \in int\left(\Delta^{T-1}\right)$ as required in the axiom. Hence,

$$H_T\left(F\right) = \sum_{t=1}^{T} \lambda_t \cdot \hat{P}_T\left(c\right) = \sum_{t=1}^{T} \lambda_t \cdot H_T\left(D_t\right)$$

Furthermore, if $F^n = D_1 \circ ... \circ D_n$ for some $n \in Z_+$ and some sets $D_1...D_n \in \mathbb{D}_T$, then

$$f_F = \frac{1}{n}\sum_{i=1}^{n} f_{D_i}.$$

Hence,

$$\begin{aligned} & \frac{\sum_{c\in C} s\left(c\right)\hat{p}_T\left(c\right) f_F\left(c\right)}{\sum_{c\in C} s\left(c\right) f_F\left(c\right)} \\ = & \sum_{i=1}^{n} \frac{\sum_{c\in C} \frac{1}{n} s\left(c\right)\hat{p}_T\left(c\right) f_{D_i}\left(c\right)}{\sum_{c\in C}\sum_{i=1}^{n} \frac{1}{n} s\left(c\right) f_{D_i}\left(c\right)} = \sum_{i=1}^{n} \frac{\sum_{c\in C} s\left(c\right)\hat{p}_T\left(c\right) f_{D_i}\left(c\right)}{\sum_{c\in C}\sum_{i=1}^{n} s\left(c\right) f_{D_i}\left(c\right)} \\ = & \sum_{i=1}^{n}\sum_{c\in C} \frac{s\left(c\right)\hat{p}_T\left(c\right)}{\sum_{c\in C}\sum_{i=1}^{n} s\left(c\right) f_{D_i}\left(c\right)} f_{D_i}\left(c\right) \\ = & \sum_{i=1}^{n} \lambda^i \sum_{c\in C} \frac{s\left(c\right)\hat{p}_T\left(c\right) f_{D_i}\left(c\right)}{\sum_{c\in C} s\left(c\right) f_{D_i}\left(c\right)} \end{aligned}$$

with

$$\lambda^i = \frac{\sum_{c\in C} s\left(c\right) f_{D_i}\left(c\right)}{\sum_{i=1}^{n}\sum_{c\in C} s\left(c\right) f_{D_i}\left(c\right)}.$$

Again, it is obvious that $\left(\lambda^i\right)_{i=1}^n \in int\left(\Delta^{n-1}\right)$ and, therefore,

$$H_T(F) = \sum_{i=1}^{n} \lambda^i H_T(D_i).$$

Since we have defined $H_T\left(c^T\right) = \hat{P}_T(c)$, and since $\hat{P}$ satisfies Linear Independence for all $T \in \{2, 3...\infty\}$, so does H. Hence, Axiom $(A3)$ also holds.

We now prove the sufficiency of the axioms for the representation.

Denote by $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ the set of rational probability vectors of dimension $|C|$. We make use of Proposition 3 from BGSS (2005, P. 1132), which we state in terms of our notation:

Proposition A.1 BGSS (2005) *Assume that* $h : \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1} \to \Delta^{|R|-1}$ *satisfies the conditions:*

(i) *for every* $f,\ f' \in \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ *and every rational* $\alpha \in (0;1)$,

$$h(\alpha f + (1-\alpha) f') = \lambda h(f) + (1-\lambda) h(f'),$$

for some $\lambda \in (0;1)$ *and*

(ii) *not all* $\{h(f)\}_{f \in \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}}$ *are collinear.*

Then there are probability vectors $(\hat{p}(c))_{c \in C} \in \Delta^{|R|-1}$ *not all of which are collinear and positive numbers* $(s(c))_{c \in C}$ *such that for every* $f \in \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$,

$$h(f) = \frac{\sum_{c \in C} s(c) f(c) \hat{p}(c)}{\sum_{c \in C} s(c) f(c)}.$$

The idea of the proof is as follows. First, we construct a sequence of sets of finite databases in such a way that the limit of this sequence is the set of infinite databases $\mathbb{D}_\infty$ with well-defined frequencies, see Lemma A.2. Hence, we can think of H_∞ as a mapping from $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ to $\Delta^{|R|-1}$. In a second step (Lemmas A.3, A.4 and Corollary 6.1), using Axioms $(A2)$ and $(A3)$, we show that H can be represented as a union of functions h, all of which satisfy properties (i) and (ii) of Proposition A.1 when restricted to $\mathbb{D}_\infty$. Next, in Lemma A.5, we apply the construction used in the proof of Proposition 3 in BGSS (2005) to determine the similarity function s for the restriction of each h to $\mathbb{D}_\infty$. Moreover, the functions h can be chosen in such a way that the similarity weights are constant across all h. The last step, Lemma A.6, consists in using Axiom $(A2)$ to show that the same similarity weights can be used for data-sets of any length $T \geq 2$.

We denote the possible frequency vectors which can be generated by a data-set of length T by:

$$Q_T = \left\{ f \in \Delta^{|C|-1} \mid f(c) = \frac{k_c}{T} \text{ for some } (k_c)_{c=1}^{|C|} \in \{0; 1...T\}^{|C|} \text{ with } \sum_{c=1}^{|C|} k_c = T \right\}.$$

Obviously, for each $T \in \{2, 3...\infty\}$, $Q_T \subseteq \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$. Our first Lemma shows that we can approximate $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ by Q_T by choosing a specific sequence of T's. We denote by $\underline{\lim}$ ($\overline{\lim}$), the inferior (superior) limit of a sequence of sets, (see BERGE (1963, P. 118) for definitions and properties).

Lemma A.2 *Consider the infinite sequence $T_1; T_2...T_m...$ with $T_m = m!$.*

$$\lim_{m\to\infty} Q_{T_m} = \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}.$$

Proof of Lemma A.2:

First, we show

$$\underline{\lim}_{m\to\infty} Q_{T_m} = \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$$

Hence, we check that for each $q \in \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$, there exists an $M \in \mathbb{Z}^+$ such that for all $m \geq M$, $q \in Q_{T_m}$. To see this, write q as a vector of ratios

$$q = \left(\frac{a_i}{b_i}\right)_{i=1}^{|C|},$$

with a_i and $b_i \in \mathbb{Z}^+$, and take the largest of the numbers b_i, $b(q) = \max_{i\in\{1...|C|\}} b_i$. Now set $M = b(q)$ and observe that for all $m \geq M$, each ratio $\frac{a_i}{b_i}$ can be written as:

$$\frac{a_i}{b_i} = \frac{a_i k_i}{b(q)!\,(b(q)+1)\,(b(q)+2)\,...m} = \frac{a_i k_i}{b_i\,(b_i-1)!\,(b_i+1)\,(b_i+2)\,...m} = \frac{a_i k_i}{T_m}$$

with

$$k_i = (b_i - 1)!\,(b_i + 1)\,(b_i + 2)\,...m.$$

Since $a_i \leq b_i$, it follows that

$$0 \leq a_i k_i \leq T_m$$

and obviously $a_i k_i \in \mathbb{Z}^+$. which proves the claim.

Second, we show that:

$$\overline{\lim}_{m\to\infty} Q_{T_m} = \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}.$$

This follows immediately from the fact that $Q_{T_m} \subset \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ for all $m \in \mathbb{Z}^+$. Hence,

$$\underline{\lim}_{m\to\infty} Q_{T_m} = \overline{\lim}_{m\to\infty} Q_{T_m} = \lim_{m\to\infty} Q_{T_m} = \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}. \blacksquare$$

The next Lemma A.3 allows us to relate the *Concatenation Axiom*, $(A2)$ (which is formulated

in terms of data-sets) to property (i) in Proposition A.1 (stated in terms of frequencies).

Lemma A.3 *Let* $\in \{2,3...\infty\}$*,* f'*,* f''*,* $f \in Q_T$ *and suppose that there is an* $\alpha \in (0;1)$ *such that:*

$$\alpha f' + (1-\alpha) f'' = f.$$

Denote by $D = (f;T)$*,* $D' = (f';T)$*,* $D'' = (f'';T)$ *the data-sets with length* T *and frequencies* f*,* f' *and* f''*. Then, there exists a* $\lambda \in (0;1)$ *such that:*

$$\lambda H(D') + (1-\lambda) H(D'') = H(D).$$

Proof of Lemma A.3:

Construct the following set of databases $D_1 = ... = D_{m-1} = D_m = D'; D_{m+1} = ... = D_n = D''$ with

$$\frac{m}{n} = \alpha.$$

Note that such integers m and n can be found as long as α is rational, which is satisfied since f, f' and $f'' \in Q_T$. Now note that:

$$\begin{aligned} D_1 \circ ... \circ D_m &= (D')^m \\ D_{m+1} \circ ... \circ D_n &= (D'')^{n-m} \\ D_1 \circ ... \circ D_n &= (D)^n, \end{aligned}$$

and, hence, by $(A2)$, there exists a vector $\mu \in int(\Delta^{n-1})$ such that:

$$\sum_{i=1}^{n} \mu_i H(D_i) = H(D).$$

Hence,

$$H(D') \sum_{i=1}^{m} \mu_i + H(D'') \sum_{i=m+1}^{n} \mu_i = H(D).$$

Setting $\lambda = \sum_{i=1}^{m} \mu_i \in (0;1)$ concludes the proof.■

We now state a lemma which shows that for every such T, we can express

$$H_T : \mathbb{D}_T \to \Delta^{|R|-1}$$

as a collection of probability functions

$$\begin{aligned} h_T &: \mathbb{D}_T \to \Delta^{|R|-1}, \\ h_T(D) &\in H_T(D) \end{aligned}$$

which satisfy properties (i) and (ii) of Proposition A.1.

Lemma A.4 *Suppose that* H_T*,* $T \in \{2,3...\infty\}$*, satisfies* $(A2)$ *and* $(A3)$*. Then, for each*

$T \in \{2, 3...\infty\}$*, there is a set of probability functions*

$$\mathcal{H}_T = \left\{h_T : \mathbb{D}_T \to \Delta^{|R|-1}\right\}$$

such that for each $T \geq 2$*,*

$$\cup_{h_T \in \mathcal{H}_T} h_T(D) = H_T(D)$$

and the following properties are satisfied:

(i') *whenever*

$$\lambda H_T(D) + (1-\lambda) H_T(D') = H_T\left(\tilde{D}\right),$$

then for each $h_T \in \mathcal{H}_T$*,*

$$\lambda h_T(D) + (1-\lambda) h_T(D') = h_T\left(\tilde{D}\right)$$

and

(ii') *not all vectors*

$$\{h_T(D)\}_{D \in \mathbb{D}_T}$$

are collinear.

Before stating the proof of Lemma A.4, we illustrate its implications by the following corollary:

Corollary 6.1 *Each* $h_T \in \mathcal{H}_T$ *as constructed in Lemma A.4 satisfies properties* (i) *and* (ii) *stated in Proposition A.1, where the set* $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ *is replaced by* Q_T *for* $T < \infty$*.*

Proof of Corollary 6.1:

For a given T, each set $D \in \mathbb{D}_T$ is uniquely identified by its frequency. Hence, property (ii') corresponds exactly to property (ii) from Proposition A.1. To see the relation between (i') and (i) recall that Lemma A.3 demonstrates that for every $T \geq 2$, every D, D', $\tilde{D} \in \mathbb{D}_T$ with frequencies f, f', $\tilde{f} \in Q_T$ (with $Q_\infty = \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$) and every rational $\alpha \in (0;1)$, such that

$$\alpha f + (1-\alpha) f' = \tilde{f},$$

we have

$$H_T\left(\tilde{D}\right) = \lambda H_T(D) + (1-\lambda) H_T(D'),$$

for some $\lambda \in (0;1)$, whereas condition (i') assures that for each $h_T \in \mathcal{H}_T$,

$$h_T\left(\tilde{D}\right) = \lambda h_T(D) + (1-\lambda) h_T(D').$$

We can now write h_T in terms of frequencies, thus obtaining the expression stated in (i):

$$h_T\left(\tilde{f}\right) = h_T(\alpha f + (1-\alpha) f) = \lambda h_T(f) + (1-\lambda) h_T(f').$$

Especially, for $\mathbb{D}_\infty$, this expression is valid for any two f and $f' \in \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ and every rational $\alpha \in (0;1)$.

Proof of Lemma A.4:

First, we show that h_T satisfying property (i') exist. By the Caratheodory Theorem, see GREEN AND HELLER (1981, P. 40), we know that for a convex set $H_T(D)$ in a finite dimensional space (such as $\Delta^{|R|-1}$), each point of the set can be represented as a convex combination of at most $|R|$ points in $\Delta^{|R|-1}$. Since we have assumed that $H_T\left((c_i)^T\right)$ are convex sets (polyhedra), we can represent each such set as:

$$H_T\left((c_i)^T\right) = \left\{ \sum_{j=1}^{|R|} \alpha_{ij}\mu_{ij} \mid \sum_{j=1}^{|R|} \alpha_{ij} = 1 \text{ and } \alpha_{ij} \geq 0 \right\},$$

where $\left(\mu_{ij}\right)_{j=1}^{|R|}$ is the above mentioned collection of points in $\mathbb{R}^{|R|-1}$. Note that since $\left((c_i)^T\right)_{i=1}^{|C|}$ is a basis of $\mathbb{D}_T$, it follows that any linear combination of data-sets (written as $(f_D; |D| = T)$) can be expressed as a linear combination of $(c_1)^T \ldots \left(c_{|C|}\right)^T$. By Lemma A.3, for every $D \in \mathbb{D}_T$,

$$H_T(D) = \sum_{i=1}^{|C|} \lambda_i H_T\left((c_i)^T\right)$$

with $\sum_{i=1}^{|C|} \lambda_i = 1$, $\lambda_i \in (0;1)$, whenever c_i occurs in D at least once and $\lambda_i = 0$, otherwise. The Caratheodory Theorem now allows us to write any such convex combination as:

$$\begin{aligned} H_T(D) &= \sum_{i=1}^{|C|} \lambda_i \left\{ \sum_{j=1}^{|R|} \alpha_{ij}\mu_{ij} \mid \sum_{j=1}^{|R|} \alpha_{ij} = 1 \text{ and } \alpha_{ij} \geq 0 \right\} = \\ &= \left\{ \sum_{i=1}^{|C|} \lambda_i \sum_{j=1}^{|R|} \alpha_{ij}\mu_{ij} \mid \sum_{j=1}^{|R|} \alpha_{ij} = 1 \text{ and } \alpha_{ij} \geq 0 \right\} \end{aligned}$$

Hence, for a fixed collection of points $\left(\mu_{ij}\right)_{i=1j=1}^{|C|\,|R|}$, we can identify each selection h_T with a vector of coefficients $(\alpha_{ij})_{i=1j=1}^{|C|\,|R|}$. Property (i') will be satisfied if we take the maximal set of such selections, i.e.

$$\Delta^{|C|\times(|R|-1)}.$$

We will now consider only functions h_T satisfying property (i') and show that it is possible to construct the set $\mathcal{H}_T$ without violating property (ii'). In terms of the representation above, property (ii') can be reformulated as follows. Suppose that for some $h_T \in \mathcal{H}_T$ (as characterized by $(\alpha_{ij})_{i=1j=1}^{|C|\,|R|}$), the vectors:

$$\left(h_T(D)\right)_{D\in\mathbb{D}_T} = \left(\sum_{j=1}^{|R|} \alpha_{ij}\mu_{ij} \right)_{i=1}^{|C|}$$

are collinear. The claim is that in the set of selections as given by $\Delta^{|C|\times(|R|-1)}$, we can find a

set of different selections, $\left(h_T^D\right)_{D\in\mathbb{D}_T}$, such that for each $\hat{D}\in\mathbb{D}_T$, $h_T^{\hat{D}}$ assumes the same values as h_T for $\hat{D}$, but is obtained by a set of vectors $\left(h_T^{\hat{D}}(D)\right)_{D\in\mathbb{D}_T}$ at least three of which are non-collinear.

Suppose first that H_T satisfies the condition of $(A3)$ for some $(c_i)^T$, $(c_j)^T$ and $(c_k)^T$, all of which are single points:

$$H(D_m)=\left\{h\left((c_m)^T\right)\right\}$$

for $m\in\{i;j;k\}$. Then, for each $\hat{h}_T(\bar{x};\bar{a})\in\mathcal{H}_T(\bar{x};\bar{a})$,

$$\begin{aligned}\hat{h}_T\left((c_i)^T\right) &= h\left((c_i)^T\right)\\ \hat{h}_T\left((c_j)^T\right) &= h\left((c_j)^T\right)\\ \hat{h}_T\left((c_k)^T\right) &= h\left((c_k)^T\right)\end{aligned}$$

must hold. Since these three vectors are not collinear by assumption, the result of the lemma obtains for this case.

Suppose, therefore that H_T satisfies the condition of $(A3)$ for some i, j and k, such that all of $H_T\left((c_m)^T\right)$ for $m\in\{i;j;k\}$ have a non-empty interior. Take some set

$$\hat{D}\in\mathbb{D}_T\backslash\left\{(c_1)^T...\left(c_{|C|}\right)^T\right\}.$$

For each $h_T\left(\hat{D}\right)\in H_T\left(\hat{D}\right)$, we have:

$$h_T\left(\hat{D}\right)=\sum_{m=1}^{|C|}\lambda_m h_T\left((c_m)^T\right)$$

for some $h_T\left((c_m)^T\right)\in H_T\left((c_m)^T\right)$. Whenever $h_T\left((c_m)^T\right)$, $m=\{i;j;k\}$ entering this representation are non-collinear for any such $h_T\left(\hat{D}\right)$, the result of the Lemma obtains. Suppose, however that $h_T\left((c_i)^T\right)$, $h_T\left((c_j)^T\right)$ and $h_T\left((c_k)^T\right)$ entering the representation are all collinear. Then, the claim of the Lemma would be violated if there is a $\hat{D}\in D_T$ such that $h_T\left(\hat{D}\right)$ can only be expressed as a linear combination of collinear vectors $h_T\left((c_m)^T\right)$. We now show that whenever it is true that $h_T\left(\hat{D}\right)$ can be expressed as a linear combination of collinear vectors, $h_T\left((c_m)^T\right)$, it can only be expressed as a linear combination of vectors $h'_T\left((c_m)^T\right)$, where:

$$h'_T\left((c_m)^T\right)=h_T\left((c_m)^T\right)$$

for all $m\neq i,j,k$, while $h'_T\left((c_m)^T\right)\in H_T\left((c_m)^T\right)$ for $m\in\{i,j,k\}$. This demonstrates that each $h_T\left(\hat{D}\right)$, can be expressed as a linear combination of non-collinear $h_T\left((c_m)^T\right)$. Taking

all admissible linear combinations of such vectors results in $h_T \in H_T$. It is obvious then that the collection of all such h_T, H_T has the required properties.

We have to consider several cases:

Case 1: For each $m \in \{i; j; k\}$,

$$h_T\left((c_m)^T\right) \in int\left(H_T\left((c_m)^T\right)\right)$$

Then, it is always possible to find ϵ_i and $\epsilon_j \in \Delta^{|R|}$ which are not-collinear to $h_T\left((c_m)^T\right)$ for $m \in \{i; j; k\}$ such that

$$h_T\left(\hat{D}\right) = \lambda_i\left(h_T\left((c_i)^T\right) + \epsilon_i\right) + \lambda_j\left(h_T\left((c_j)^T\right) + \epsilon_j\right) + \sum_{\substack{m=1 \\ m\neq i;j}}^{|C|} \lambda_m h_T\left((c_m)^T\right), \quad \text{(A-1)}$$

and

$$\lambda_i\epsilon_i + \lambda_j\epsilon_j = 0.$$

Defining $h'_T\left((c_m)^T\right) = h_T\left((c_m)^T\right) + \epsilon_m$ with $\epsilon_m = 0$ for $m \notin \{i, j\}$ gives the desired result.

Case 2: Let $h_T\left((c_m)^T\right)$ be extreme points of $\left(H_T\left((c_m)^T\right)\right)$ for every $m \in \{i; j; k\}$. Then, Axiom $(A3)$ insures that not all of these points are collinear and, hence, the result of the lemma obtains.

Case 3: Let $h_T\left((c_m)^T\right) \in bd\left(H_T\left((c_m)^T\right)\right)$, but not extreme points for all $m \in \{i; j; k\}$.

Case 3a: Suppose first that the hyperplanes containing the sides of the polyhedra on which $h_T\left((c_i)^T\right)$ and $h_T\left((c_j)^T\right)$ lie are not parallel. In that case, it is obvious that there exist ϵ_i such that

$$h_T\left((c_i)^T\right) + \epsilon_i \in int\left(H_T\left((c_i)^T\right)\right)$$

and ϵ_j such that

$$h_T\left((c_j)^T\right) + \epsilon_j \in bd\left(H_T\left((c_j)^T\right)\right)$$

so that:

$$\lambda_i\epsilon_i + \lambda_j\epsilon_j = 0$$

and, hence, the equality in A-1 obtains. (This can be done, e.g. by choosing ϵ_j to lie in the same hyperplane as $h_T\left((c_j)^T\right)$ and choosing $\left(\epsilon_i; h_T\left((c_i)^T\right)\right)$ to be parallel to the hyperplane on which $h_T\left((c_j)^T\right)$ lies. An ϵ_i in the interior of $H_T\left((c_j)^T\right)$ exists by the assumption that the two hyperplanes are not parallel).

Case 3b: Suppose now that all three of the hyperplanes containing the sides of the polyhedra on which $h_T\left((c_m)^T\right)$ lie are parallel, but at least two of them are distinct. Then choose vectors ϵ_i

and ϵ_j such that

$$\lambda_i \epsilon_i + \lambda_j \epsilon_j = 0$$

and both ϵ_i and ϵ_j are parallel to the hyperplanes containing the sides of the polyhedra on which $h_T\left(\left(c_m\right)^T\right)$ lie. It is obvious that ϵ_i and ϵ_j can always be chosen in such a way that

$$h_T\left(\left(c_i\right)^T\right) + \epsilon_i, h_T\left(\left(c_j\right)^T\right) + \epsilon_j \text{ and } h_T\left(\left(c_k\right)^T\right)$$

are not collinear.

Case 3c: If the three hyperplanes containing the sides of the polyhedra on which $h_T\left(\left(c_m\right)^T\right)$ lie, coincide, there are two possibilities: either at least one of the points belongs to the interior of a face on this hyperplane or all of the points lie on edges of the polyhedra. Let $h_T\left(\left(c_i\right)^T\right)$ belong to the interior of a face. If the edge containing, say $h_T\left(\left(c_j\right)^T\right)$ is not collinear to the edge containing $h_T\left(\left(c_k\right)^T\right)$, then, it is obviously possible to find ϵ_i and ϵ_j satisfying the necessary condition A-1. The idea is to move $h_T\left(\left(c_j\right)^T\right)$by ϵ_j along the edge to which it belongs, while moving the interior point $h_T\left(\left(c_i\right)^T\right)$ in the opposite direction by the use of ϵ_i. If both edges are collinear, then ϵ_j can be chosen in such a way so as to move $h_T\left(\left(c_j\right)^T\right)$ into the interior of $H_T\left(\left(c_j\right)^T\right)$, whereas again it is always possible to move the interior point $h_T\left(\left(c_i\right)^T\right)$ into the exactly opposite direction by means of ϵ_j.

If at least two of the edges are not parallel, then the existence of ϵ_i and ϵ_j is obvious, as in the case of non-parallel hyperplanes. If the edges are parallel but distinct lines in this hyperplane, proceed as in the case of three parallel but distinct hyperplanes. If all of the lines containing the edges coincide, then all vertices contained in these edges must be collinear, which is excluded by $(A3)$.■

Lemma A.5 *Let $D \in \mathbb{D}_\infty$. Then,*

$$H_\infty(D) = \left\{ \frac{\sum_{c \in C} s(c)\, \hat{p}_\infty(c)\, f_D(c)}{\sum_{c \in C} s(c)\, f_D(c)} \mid \hat{p}_\infty(c) \in \hat{P}_\infty(c) \right\},$$

where

$$\hat{P}_\infty(c) = H\left((c)^\infty\right),$$

(and hence, satisfy Linear Independence*) and $s(c)$ are given by the unique (up to a multiplication by a positive number) solution of the equation:*

$$\frac{\sum_{i=1}^{|C|} \frac{1}{|C|} s(c_i)\, \hat{p}_\infty(c_i)}{\sum_{i=1}^{|C|} \frac{1}{|C|} s(c_i)} = \sum_{i=1}^{|C|} \lambda_i h\left((c_i)^\infty\right).$$

Proof of Lemma A.5:

Apply the construction in Lemma A.4, to the sequence T_m (as defined in Lemma A.2). Letting $m \to \infty$, we can represent H_∞ as a selection of functions h_∞ which satisfy all of the conditions of Proposition A.1. We can, therefore, apply directly the result of the proposition and state, for each h_∞, the existence of unique vectors

$$\hat{p}_\infty(c_1) ... \hat{p}_\infty\left(c_{|C|}\right)$$

such that

$$h_\infty\left((c_i)^\infty\right) = \frac{\sum_{c\in C} s(c)\,\hat{p}_\infty(c)\,f_{(c_i)^\infty}(c)}{\sum_{c\in C} s(c)\,f_{(c_i)^\infty}(c)} = \hat{p}_\infty(c_i),$$

or

$$\hat{p}_\infty(c_1) = h\left((c_1)^\infty\right) ... \hat{p}_\infty\left(c_{|C|}\right) = h\left(c_{|C|}\right)^\infty.$$

Taking the union of all such vectors $\hat{p}$, we thus obtain the sets

$$\hat{P}_\infty(c_i) = \cup_{h_\infty\in\mathcal{H}_\infty} h_\infty(c_i)^\infty = H_\infty\left((c_i)^\infty\right) \text{ for } i \in \{1...|C|\}.$$

These sets trivially satisfy the conditions of Axiom $(A3)$. We can now determine the similarity function for each of the vectors

$$\hat{p}^1_\infty(c_1) ... \hat{p}_\infty\left(c_{|C|}\right)$$

separately by solving:

$$\frac{\sum_{i=1}^{|C|} \frac{1}{|C|} s(c_i)\,\hat{p}_\infty(c_i)}{\sum_{i=1}^{|C|} \frac{1}{|C|} s(c_i)} = \sum_{i=1}^{|C|} \lambda_i h_\infty\left((c_i)^\infty\right). \tag{A-2}$$

For the case $|C| = 3$, the condition that $h\left((c_1)^\infty\right)$, $h\left((c_2)^\infty\right)$ and $h\left((c_3)^\infty\right)$ are non-collinear implies that this system has a unique solution, $\{s_\infty((c_i))\}_{i=1}^3$. For the case of $|C| > 3$, we can apply Step 2 of the proof of BGSS (2005), which implies that no matter which three non-collinear vectors are chosen, the resulting similarity functions differ only with respect to a multiplication by a positive number. Lemma A.4 insures that $(\lambda_i)_{i=1}^{|C|}$ remain the same for all functions h. Since $\hat{p}_\infty(c_i) = h_\infty\left((c_i)^\infty\right)$ it follows that the unique (up to a multiplication by a positive number) solution to this equation does not depend on the chosen vector and is given by:

$$s(c_i) = \lambda_i. \blacksquare$$

Lemma A.6 *For every* $T \in \{2, 3...\infty\}$,

$$H_T(D) = \left\{ \frac{\sum_{c\in C} s(c)\,\hat{p}_T(c)\,f_D(c)}{\sum_{c\in C} s(c)\,f_D(c)} \mid \hat{p}_T(c) \in \hat{P}_T(c) \right\},$$

where

$$\hat{P}_T(c) = H\left((c)^T\right),$$

and $s(c)$ *are the unique (up to a multiplication by a positive number) solution of equation A-2.*

Proof of Lemma A.6:

First note that using the argument in the proof of Proposition 3 in BGSS (2005, P. 1134) we can show that the solution of the system:

$$\begin{aligned} & \frac{\frac{T-1}{T}s(c_1)\hat{p}_\infty(c_1) + \frac{1}{T}s(c_2)\hat{p}_\infty(c_2)}{\frac{T-1}{T}s(c_1) + \frac{1}{T}s(c_2)} \\ = \; & \lambda^1 h_\infty\left((c_1)^\infty\right) + \left(1-\lambda^1\right)h_\infty\left((c_2)^\infty\right) \\ & \dots \\ & \frac{\frac{T-1}{T}s\left(c_{|C|-1}\right)\hat{p}_\infty\left(c_{|C|-1}\right) + \frac{1}{T}s\left(c_{|C|}\right)\hat{p}_\infty\left(c_{|C|}\right)}{\frac{T-1}{T}s\left(c_{|C|-1}\right) + \frac{1}{T}s\left(c_{|C|}\right)} \\ = \; & \lambda^{|C|-1}h_\infty\left(\left(c_{|C|-1}\right)^\infty\right) + \left(1-\lambda_1^{|C|-1}\right)h_\infty\left(\left(c_{|C|}\right)^\infty\right) \end{aligned} \tag{A-3}$$

is identical (up to a multiplication by a positive number) to the solution of equation A-2. Note that this argument uses only properties (i) and (ii), but does not make use of the fact that h_∞ is defined on the set $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$.

Let $T < \infty$. Corollary 6.1 shows that properties (i) and (ii) stated in Proposition A.1 are satisfied for all finite data-sets with equal length T as long as the set of possible values of f and f' is restricted to Q_T.

Observe that for each selection h_T, we have:

$$h_T\left((c_i)^T\right) = \frac{\sum_{c\in C} s(c)\hat{p}_T(c) f_{(c_i)^T}(c)}{\sum_{c\in C} s(c) f_{(c_i)^T}(c)} = \hat{p}_T(c_i)$$

and define

$$\hat{P}_T(c_i) = H_T\left((c_i)^T\right)$$

Note that, for i and $j \in \{1...|C|\}$we can write:

$$\left(\underbrace{c_i...c_i}_{T-1\text{-times}}; c_j\right)^T = \left((c_i)^T\right)^{T-1} \circ (c_j)^T$$

and conclude, by Axiom $(A2)$ and Lemma A.3 that

$$H_T\left(\underbrace{c_i...c_i}_{(T-1)\text{-times}}; c_j\right) = \lambda H_T\left((c_i)^T\right) + (1-\lambda) H_T\left((c_j)^T\right).$$

for some $\lambda \in (0;1)$. Lemma A.4 shows that the same values of λ can be used for each selection

h_T of H_T. Axiom $(A2)$ guarantees that for any $k \in \mathbb{Z}_+$,

$$\left(\underbrace{c_i ... c_i}_{T-1\text{-times}} ; c_j \right)^{kT} = \left((c_i)^T \right)^{k(T-1)} \circ (c_j)^{kT}$$

implies

$$H_{kT} \left(\underbrace{c_i ... c_i}_{k(T-1)\text{-times}} ; \underbrace{c_j ... c_j}_{k\text{-times}} \right) = \lambda H_T \left((c_i)^T \right) + (1-\lambda) H_T \left((c_j)^T \right).$$

Letting $k = T_m = m!$ and $m \to \infty$, we get:

$$\lim_{T_m \to \infty} H \left(\underbrace{c_i ... c_i}_{T_m(T-1)\text{-times}} ; \underbrace{c_j ... c_j}_{T_m\text{-times}} \right) = \lambda H_\infty \left((c_i)^\infty \right) + (1-\lambda) H_\infty \left((c_j)^\infty \right)$$

and from Lemma A.5, we know that:

$$\lambda h_\infty \left((c_i)^\infty \right) + (1-\lambda) h_\infty \left((c_j)^\infty \right) = \frac{\frac{T-1}{T} s(c_i) \hat{p}_\infty(c_i) + \frac{1}{T} s(c_j) \hat{p}_\infty(c_j)}{\frac{T-1}{T} s(c_i) + \frac{1}{T} s(c_j)}$$

for each selection $h_\infty \in H_\infty$.

Hence, we can determine the similarity function for data-sets of length T by solving the system of equations:

$$\begin{aligned} & \frac{\frac{T-1}{T} s(c_1) \hat{p}_T(c_1) + \frac{1}{T} s(c_2) \hat{p}_T(c_2)}{\frac{T-1}{T} s(c_1) + \frac{1}{T} s(c_2)} \\ = \; & \lambda^1 h \left((c_1)^T \right) + \left(1 - \lambda^1\right) h \left((c_2)^T \right) \\ & \dots \\ & \frac{\frac{T-1}{T} s\left(c_{|C|-1}\right) \hat{p}_T\left(c_{|C|-1}\right) + \frac{1}{T} s\left(c_{|C|}\right) \hat{p}_T\left(c_{|C|}\right)}{\frac{T-1}{T} s\left(c_{|C|-1}\right) + \frac{1}{T} s\left(c_{|C|}\right)} \\ = \; & \lambda^{|C|-1} h \left(\left(c_{|C|-1}\right)^T \right) + \left(1 - \lambda^{|C|-1}\right) \left(h \left(c_{|C|}\right)^T \right) \end{aligned}$$

in which the λ-values are identical to those in equation A-3 above. Since the selections h_T satisfy properties (i) and (ii) of Proposition A.1 restricted to Q_T and since the argument from the proof of Proposition 3 in BGSS (2005) used above does not depend on the set Q_T, we can use it again to claim that the unique solution to this system coincides with the solution of A-3 and is also independent of the values of $\hat{p}_T(c)$ as long as

$$\hat{p}_T(c_i) = h \left((c_i)^T \right)$$

holds. Hence, we can use the similarity function determined for $\mathbb{D}_\infty$, for any $\mathbb{D}_T$ with $T < \infty$.■

Proof of Lemma 5.1:

Suppose that the frequency of r in a data-set $D = \left\{ (x; a; r_t)_{t=1}^T \right\}$ is given by $f_D(r)$. Consider

the sequence of data-sets D^k as $k \to \infty$ and note that by $(A4)$, as $k \to \infty$,

$$H\left(D^{\infty}\right) \to \left\{h\left(D^{\infty}\right)\right\} = \left\{f_{D^k}\left(r\right)\right\} = \left\{f_{D^{\infty}}\left(r\right)\right\}.$$

By $(A5)$, for each k,

$$H\left(D^k\right) \subset H\left(D^{k-1}\right).$$

Hence, for each k, there is an $h \in H\left(D^k\right)$ such that

$$h_r\left(D\right) = f_{D^k}\left(r\right).$$

Especially, for $k = 1$, there is an $h \in H\left(D\right)$ such that

$$h_r\left(D\right) = f_D\left(r\right). \blacksquare$$

Proof of Theorem 5.2:

To simplify notation, for this proof, we let $c_i^{x;a}$ denote a case $(x; a; r_i)$, where only the outcome r varies while x and a remain fixed. To see that the proposition holds note that we construct the elements of $\hat{P}_T\left(c_i^{x;a} \mid x; a\right)$ by using only the data-set $\left(c_i^{x;a}\right)^T$ and setting for each selection h,

$$\hat{p}_T\left(c_i^{x;a} \mid x; a\right) =: h\left(\left(c_i^{x;a}\right)^T \mid x; a\right).$$

Hence,

$$\hat{P}_T\left(c_i^{x;a} \mid x; a\right) = H\left(\left(c_i^{x;a}\right)^T \mid x; a\right).$$

$(A5)$, *Accumulation of knowledge* ascertains that

$$H_{T+1}\left(\left(c_i^{x;a}\right)^{T+1} \mid x; a\right) \subset H_T\left(\left(c_i^{x;a}\right)^T \mid x; a\right).$$

Now note that, if Axiom $(A4)$, Learnability, holds, we know that for $c_i^{x;a} = (x; a; r^i)$

$$\lim_{T\to\infty} H\left(\left(c_i^{x;a}\right)^T \mid x; a\right) = f_{D_{\infty}^i} = \left(0; 0...0; \underbrace{1}_{i^{\text{th}}\text{-position}}; 0...0\right) = \lim_{T\to\infty} \hat{P}_T^i\left(\left(c_i^{x;a}\right) \mid x; a\right).$$

The inclusion property shown above ascertains that

$$\left(0; 0...0; \underbrace{1}_{i^{\text{th}}\text{-position}}; 0...0\right) \in \hat{P}_T\left(c_i^{x;a} \mid x; a\right)$$

for every T. Now consider all cases $\left(x; a; r\right)_{r\in R}$ and the data-sets $\left(x; a; r\right)^T$. For any two such

sets with outcomes r_i and r_j, we know that

$$\begin{aligned}
\lim_{T\to\infty} H\left((x;a;r_i)^T \circ (x;a;r_j)^T \mid x;a\right) &= \lim_{T\to\infty} f_{(x;a;r_i)^T\circ(x;a;r_j)^T} = \\
&= \lim_{T\to\infty} \frac{1}{2} H\left((x;a;r_i)^T \mid x;a\right) + \\
+\frac{1}{2} H\left((x;a;r_j)^T \mid x;a\right) &= \\
&= \frac{1}{2} f_{(x;a;r_i)^T} + \frac{1}{2} f_{(x;a;r_j)^T} = \\
&= \left(0;0...\underbrace{\frac{1}{2}}_{i^{\text{th}}\text{ position}};0...\underbrace{\frac{1}{2}}_{j^{\text{th}}\text{ position}};0...0\right)
\end{aligned}$$

Now, expressing

$$H\left((x;a;r_i)^T \circ (x;a;r_j)^T \mid x;a\right)$$

in terms of similarity gives:

$$\begin{aligned}
&\left(0;0...\underbrace{\frac{1}{2}}_{i^{\text{th}}\text{ position}};0...\underbrace{\frac{1}{2}}_{j^{\text{th}}\text{ position}};0...0\right) \\
= &\frac{\frac{1}{2}s\left((x;a);(x;a;r_i)\right)e_i + \frac{1}{2}s\left((x;a);(x;a;r_j)\right)e_j}{\frac{1}{2}s\left((x;a);(x;a;r_i)\right) + \frac{1}{2}s\left((x;a);(x;a;r_i)\right)},
\end{aligned}$$

which implies

$$s\left((x;a);(x;a;r_i)\right) = s\left((x;a);(x;a;r_j)\right),$$

for all r_i, $r_j \in R$, which after normalization can be written as:

$$s\left((x;a);(x;a;r)\right) = 1$$

for all $r \in R$. ■

References

AHN, D. (2008). "Ambiguity Without a State Space", *Review of Economic Studies 71*, 3-28.

BERGE, C. (1963). *Topological Spaces*, The Macmillan Company, New York.

BILLOT, A., GILBOA, I., SAMET, D. AND SCHMEIDLER, D. (2005). "Probabilities as Similarity-Weighted Frequencies". *Econometrica 73*, 1125-1136.

CHATEAUNEUF, A., EICHBERGER , J., AND GRANT, S. (2007). "Choice Under Uncertainty with the Best and the Worst in Mind: Neo-Additive Capacities", *Journal of Economic Theory 137,* 538-567

COIGNARD, Y., AND JAFFRAY, J.-Y. (1994). "Direct Decision Making" in: *Decision Theory and Decision Analysis: Trends and Challenges,* Rios, S. (ed.). Boston: Kluwer Academic Publishers.

DURRETT, R. (2005). *Probability: Theory and Examples*, Thomson Brooks / Cole, Australia.

EICHBERGER, J. AND GUERDJIKOVA, A. (2008). "From Cases to Lotteries — Learning from Data with the Best and the Worst in Mind"*, mimeo,* Cornell University.

EPSTEIN, L. AND SCHNEIDER, M. (2007). "Learning Under Ambiguity", *Review of Economic Studies 74,* 1275-1303.

GAJDOS, TH., HAYASHI, T., TALLON, AND J.-M., VERGNAUD, J.-C. (2007). "Attitude Towards Imprecise Information", *fortchcoming in Journal of Economic Theory.*

GAYER, G. (2007). "Perception of Probabilities in Situations of Risk A Case Based Approach"*, mimeo,* University of Haifa.

GHIRARDATO, P., MACCHERONI, F., AND MARINACCI, M. (2004). "Differentiating Ambiguity abnd Ambiguity Attitude", *Journal of Economic Theory 118*, 133-173.

GILBOA, I., LIEBERMAN, O. AND SCHMEIDLER, D. (2004). "Empirical Similarity", *Review of Economics and Statistics,* forthc.

GILBOA, I., AND SCHMEIDLER, D. (2001). *A Theory of Case-Based Decisions.* Cambridge, UK: Cambridge University Press.

GILBOA, I., AND SCHMEIDLER, D. (1997). "Act Similarity in Case-Based Decision Theory", *Economic Theory* 9: 47-61.

GILBOA, I., AND SCHMEIDLER, D. (1989). "Maxmin Expected Utility with a Non-Unique Prior", *Journal of Mathematical Economics* 18, 141-153.

GILBOA, I., SCHMEIDLER, D., WAKKER, P. (2002). "Utility in Case-Based Decision Theory", *Journal of Economic Theory 105,* 483-502.

GREEN, J AND HELLER, W. P. (1981). "Mathematical Analysis and Convexity" in: *Handbook of Mathematical Economics*, Arrow, K. J. and Intriligator, M. D. (eds.). Amsterdam: North-Holland Publishing Company.

GONZALES, CH. AND JAFFRAY, J.-Y. (1998). "Imprecise Sampling and Direct Decision Making", *Annals of Operations Research 80*, 207-235.

GUERDJIKOVA, A. (2007). "Preference for Diversification with Similarity Considerations", in: *Uncertainty and Risk. Mental, Formal, Experimental Representations,* Abdellaoui, M., Luce, R. D., Machina, M. J., Munier, B. (eds.). Berlin: Springer.

GUERDJIKOVA, A. (2008). "Case-Based Learning with Different Similarity Functions", *Games and Economic Behavior 63*, 107-132.

KEYNES, J. M. (1921). *A Treatise on Probability*. London:Macmillan.

KLIBANOFF, P., MARINACCI, M., AND MUKERJI, S. (2005). "A Smooth Model of Decision Making Under Uncertainty", *Econometrica 73*, 1849-1892.

LUCE, B.R. AND O'HARA, A. (2003). *A Primer on Bayesian Statistics in Health Ecionomics and Operations Research.* MEDTAP International.

NEHRING, K. (1999). "Diversity and the Geometry of Similarity", University of California at Davis, *mimeo*.

NEHRING, K. AND PUPPE, C. (2002). "A Theory of Diversity", *Econometrica 70,* 1155-1198.

PESKI, M. (2007). "Learning through Patterns", *mimeo,* University of Chicago.

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

STINCHCOMBE, M. (2003). "Choice and Games with Ambiguity as Sets of Probabilities", *Working paper,* University of Texas, Austin.