

Multiple Priors as Similarity Weighted Frequencies¹

Jürgen Eichberger² and Ani Guerdjikova³

This version: August 31, 2007

Abstract

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 entertains a set of priors expressing his hypotheses 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 and Marcin Peski, as well as the participants of the Conference on Risk, Uncertainty and Decisions in Tel Aviv 2007 and seminar participants at Heidelberg, Cergy-Pontoise and Cornell for helpful comments and suggestions. Financial support from the DFG and from the Center for Analytic Economics at Cornell is gratefully acknowledged.

² 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 does 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 is 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 represent an ideal method of data collection. In this case, decision makers can aggregate information directly in a probability distribution over unknown states. Indeed, in this context, it is of little consequence whether one follows a frequentist or a Bayesian approach. Both approaches will lead to the same probability distribution as more evidence becomes available with an increasing database.

In most real-life decision problems, however, decision makers do not have available data derived from explicitly designed experiments. 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 distribution. 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 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 data base 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 content and size of the data-set. In a second step, we assume that the confidence of the decision maker increases as data accumulates and that the set of probability distributions converges to the observed frequency when the data-set becomes large.

As in BGSS (2005), the question remains open which decision criterion one should use given the information contained in a set of multiple priors. 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).

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 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 allow 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 inhomogeneous 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. An example in Section 5 illustrates the properties of this representation. Section 6 deals with the special case of data collected in controlled experiments and Section 7 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 $t = 1 \dots T$:

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

The set of all data-sets, denoted by $\mathbb{D} := \bigcup_{T \geq 1} C^T$, is the set of databases of any length $T \in \mathbb{Z}^+$. 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 $x_0 \in X$. We will assume that the decision maker associates a set of probability distributions over outcomes R ,

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

with the action a_0 in the situation characterized by x_0 given the database $D \in \mathbb{D}$. Formally, $H : \mathbb{D} \times X \times A \rightarrow \Delta^{|R|-1}$ is a correspondence which maps $\mathbb{D} \times X \times A$ into *compact* and *convex subsets* of $\Delta^{|R|-1}$.

We interpret $H(D | x_0; a_0)$ as the set of probability distributions over outcomes which the decision maker takes into consideration given the database D . In applications one may think of this set of probabilities as a neighborhood of the frequencies of relevant observations in D . With such applications in mind, the assumption that $H(D | x_0; a_0)$ is a compact and convex subset of $\Delta^{|R|-1}$ appears reasonable. Elements of this set are denoted by $h(D | x_0; a_0)$ and we write $h_r(D | x_0; a_0)$ for the probability assigned to outcome r by the probability distribution

$h(D \mid x_0; a_0)$.

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 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 observed⁴ in the past may serve the physician in this assessment of probabilities over outcomes.

A case is a combinations 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 . ■

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

⁴ 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.

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 consumer 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 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.,

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

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 who had been successfully offered a movie $a_t \in A$. Given this database D the system assesses the likelihood $H(D | x_0; a_0)$ of the customer x_0 renting a movie from the suggested category a_0 . The set of probability distributions over R , $H(D | x_0; a_0)$, 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 $H(D | x_0; a_0)$ large. Alternatively, if there are many observations for a returning customer in the database, the set $H(D | x_0; a_0)$ 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 black and white balls in 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

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

Suppose that, after T rounds, the decision-maker has a database $D = ((1; B); (3; W); \dots; (2; B)) \in C^T$.

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 payouts $\{1; 0\}$ 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 $H(D | 2; a_0)$. 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.,

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

The set of probabilities over outcomes $H(D | 2; a_0)$ 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 2⁵. 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 ,

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

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 $T \geq 2$ 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 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 $(x_0; a_0)$. $\hat{P}_T(c \mid x_0; a_0)$ is a set of probability distributions over outcomes and represents the beliefs entertained by the decision maker when the database $D = \underbrace{(c \dots c)}_{T\text{-times}}$ is observed. The axioms suggested below will imply (up to a normalization) unique similarity weights $s(c \mid x_0; a_0)$, which do not depend on T and, for each $T \geq 2$, 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

⁵ Part III of KEYNES (1921) provides an extensive review of the literature on induction from cases to probabilities.

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 becomes available conveys important information. We will abstract here from this possibility and assume that only the data matters for the probability distributions over outcomes.

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

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

According to Axiom A1 the order of cases in a database $D = (c_t)_{t=1}^T$ is irrelevant. The set of probability distributions over outcomes is invariant with respect to the sequence in which data arrives. 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.

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

Definition 4.1 (Concatenation) For 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' .

The following notational conventions are useful.

Notation By Axiom A1 a concatenation is a commutative operation on databases. Hence, we will write $D^k = \underbrace{D \circ \dots \circ D}_{k\text{-times}}$ for k concatenations of the same database D . In particular, a database consisting of k -times the same case c can be written as $(c)^k$. ■

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

The *Concatenation Axiom* of BGSS (2005) implies that, for any k , the databases D and D^k map into the same probability distribution 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 ; $h((c)^{10000}) = e_r$. Hence, e_r should be in the set of probability distributions $H((c)^{10000})$ associated with the database $(c)^{10000}$. 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., $\bar{h}(D)$ with $\bar{h}_{r'}(D) = \frac{1}{|R|}$ for all $r' \in R$. It appears perfectly reasonable to include \bar{h} in $H((c))$ but not in $H((c)^{10000})$.

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.

Denote by $\mathbb{D}_T := C^T$ the set of databases of length T . Recall that the convex combination of two sets H and H' is defined by

$$\lambda H + (1 - \lambda) H' = \{\lambda h + (1 - \lambda) h' \mid h \in H \text{ and } h' \in H'\}.$$

Axiom A2 (*Concatenation*) Consider a data set $F \in \mathbb{D}_T$ and, for some $n \in \mathbb{Z}_+$, let $D_1 \dots D_n \in \mathbb{D}_T$ be such that $D_1 \circ \dots \circ D_n = F^n$. Then, there exists a vector $(\lambda_1 \dots \lambda_{n-1}) \in \text{int}(\Delta^{n-1})$

such that, for every $k \in \mathbb{Z}^+$,

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

To illustrate the axiom, consider an example. Let $T = 3$, $|C| = 3$ and take

$$F = \left(f = \left(\frac{1}{3}; \frac{2}{3}; 0 \right); T = 3 \right).$$

Note that F^3 can be represented as concatenation of $D_1 = (c_1)^3$ and $D_2 = D_3 = (c_2)^3$:

$$F^3 = D_1 \circ D_2 \circ D_3 = D_1 \circ D_2^2.$$

Axiom A2 asserts the existence of $\lambda_1, \lambda_2, \lambda_3 \in (0; 1)$ such that $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and:

$$\lambda_1 H(D_1) + \lambda_2 H(D_2) + \lambda_3 H(D_3) = H(F).$$

Each data-set with at least two observations can be represented as a concatenation by choosing n and the data-sets $(D_i)_{i=1}^n$ in an appropriate way, e.g., by choosing the basis of the vector space \mathbb{D}_T , with $D_i = (c_i)^T$ and that the representation is in general non-unique.

In spirit, Axiom A2 is very similar to the *Concatenation Axiom* introduced by BGSS (2005). The main difference is that we restrict the axiom to data sets of equal length. Restricted to the set \mathbb{D}_T , Axiom A2 has the following implication: if the evidence of each of the n data-sets of equal length $D_1 \dots D_n$ suggests that a given outcome r is possible, i.e. for all $i \in \{1 \dots n\}$,

$$h_r(D_i) > 0 \text{ for some } h \in H(D_i)$$

then r must be considered possible under the data-set F , i.e.

$$h_r(F) > 0 \text{ for some } h \in H(F)$$

resulting from the concatenation of these data-sets, while controlling for the length of the data-set.

The restriction to sets of equal length is important for our approach since databases of different length may give rise to different degrees of confidence. To see this, consider the databases D and $D^2 = D \circ D$. Since the database D^2 contains twice the number of cases in database D , it appears reasonable to assume that the decision maker should be more confident to make a prediction based on the bigger database D^2 than on D . In other words, it might be that D does

not contain enough observations to exclude the possibility of a given outcome r , i.e. $h(r) > 0$ for some $h \in H(D)$, whereas the data-set D^2 is sufficiently long to imply $h_r(D^2) = 0$ for all $h \in H(D^2)$. Applying the *Concatenation Axiom* of BGSS (2005), we would conclude that for some $\lambda \in (0, 1)$

$$H(D^2) = H(D \circ D) = \lambda H(D) + (1 - \lambda)H(D) = H(D),$$

which seems counterintuitive in this context. 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 $\lambda H(D) + (1 - \lambda)H(D) = H(D)$, which is trivially satisfied for any set D .

Remark 4.1 *Note that Axiom A2 requires the weights λ to be constant across different T 's, as long as the frequencies of the data-sets entering the concatenation remain unchanged. This assumption is crucial for our result that the similarity function is uniquely determined. In particular, we can construct the similarity function for a specific class of data-sets containing infinitely many observations by using the methods of BGSS (2005) and then, using this assumption, extend the representation to all finite data-sets. ■*

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

Axiom A3 (*Linear Independence*) For every $T \in \mathbb{Z}^+$, the basis of \mathbb{D}_T , $(c_1)^T, \dots, (c_{|C|})^T$ satisfies the following condition:

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

– either singletons

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

and $h((c_i)^T)$, $h((c_j)^T)$ and $h((c_k)^T)$ 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 $H((c_i)^T)$ 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.

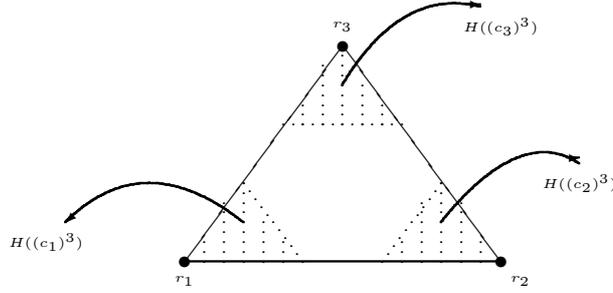


Figure 1. Non-collinear sets of priors

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

Theorem 4.1 *Let H be a correspondence $H : \mathbb{D} \rightarrow \Delta^{|R|-1}$ 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) for every $T \geq 2$.*
- (ii) *There exists a function*

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

and, for each $T \in \mathbb{Z}^+$, $T \geq 2$, there exists a correspondence

$$\hat{P}_T : C \rightarrow \Delta^{|R|-1}$$

satisfying Linear Independence and such that for any $D \in \mathbb{D}_T$

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

Moreover, for each T , \hat{P}_T 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 $T < \infty$.

5 Discussion of the Main Result

We will use Example 3.1 to illustrate the representation derived in Theorem 4.1 and to discuss the notion of similarity.

5.1 Leading Example

Example 3.1 resumed

Consider a doctor who has to choose one of two treatments, $a \in A = \{a_1; a_2\}$. In past treatments, one has recorded only two characteristics of patients, high blood pressure, x_h , or low blood pressure, x_l . Hence, the set of potentially case-relevant data comprises $x \in X = \{x_h; x_l\}$. Finally, three outcomes of the treatment have been registered, say r_1 , success, r_2 , no effect, and r_3 , failure, i.e., $R = \{r_1; r_2; r_3\}$.

In this situation, databases D of any length T will be made up of the following twelve cases:

$$C = \left\{ \begin{array}{l} \overline{c_1 = (x_1; a_1; r_1)} \\ \overline{c_2 = (x_1; a_1; r_2)} \\ \overline{c_3 = (x_1; a_1; r_3)} \\ \overline{c_4 = (x_1; a_2; r_1)} \\ \overline{c_5 = (x_1; a_2; r_2)} \\ \overline{c_6 = (x_1; a_2; r_3)} \\ \overline{c_7 = (x_2; a_1; r_1)} \\ \overline{c_8 = (x_2; a_1; r_2)} \\ \overline{c_9 = (x_2; a_1; r_3)} \\ \overline{c_{10} = (x_2; a_2; r_1)} \\ \overline{c_{11} = (x_2; a_2; r_2)} \\ \overline{c_{12} = (x_2; a_2; r_3)} \end{array} \right\}$$

For each T and each case c_i , $i = 1 \dots 12$, consider the following sets of probabilities over out-

comes:

$$\begin{aligned}\hat{P}_T(c_i) &:= \{\hat{p} \in \Delta^2 \mid \hat{p}_1 \geq (1 - \frac{\varepsilon}{T})\} & \text{for } i = 1; 4; 7; 10, \\ \hat{P}_T(c_i) &:= \{\hat{p} \in \Delta^2 \mid \hat{p}_2 \geq (1 - \frac{\varepsilon}{T})\} & \text{for } i = 2; 5; 8; 11, \\ \hat{P}_T(c_i) &:= \{\hat{p} \in \Delta^2 \mid \hat{p}_3 \geq (1 - \frac{\varepsilon}{T})\} & \text{for } i = 3; 6; 9; 12\end{aligned}$$

for some $\varepsilon > 0$.

This assignment of probabilities over outcomes can be given the following interpretation. Databases with constant $(x; a)$ provide a controlled experiment about the probabilities over outcomes, e.g., a database made up only of cases c_1 , c_2 and c_3 would generate a frequency on R which might serve as an estimate for the probabilities on R . If there is some ambiguity ε about this estimate, which appears natural if there are few observations, one may assume that this ambiguity decreases as the number of confirming observations T grows. In particular, databases containing only a single case $(x; a; r)$ may have the probability distribution yielding the outcome r with probability 1 as a natural first estimate. Note also that the $\hat{P}_T(c_i)$ satisfy Linear independence (Axiom A3).

We will assume fixed similarity weights $(s_1; \dots; s_{12})$ for the twelve basic cases. For an arbitrary database D of length T with frequencies of cases f_D one obtains the following set $H_T(D)$ of probability distributions:

$$\begin{aligned}H_T(D) \quad : \quad &= \left\{ h \in \Delta^2 \mid h = \left(\sum_{i=1}^{12} s(c_i) f_D(c_i) \right)^{-1} \left(\sum_{i=1}^{12} s(c_i) f_D(c_i) \hat{p}_T(c_i) \right), \right. \\ &\left. \hat{p}_T(c_i) \in \hat{P}_T(c_i), i \in \{1; \dots; 12\} \right\}.\end{aligned}$$

5.2 Similarity

The similarity weights $s(c_i)$ of Theorem 4.1 have to be seen in relation to the decision situation $(x_0; a_0)$ under consideration. The notation $s(c \mid x_0; a_0)$ emphasizes this relationship. If a decision situation $(x_0; a_0)$ is part of the cases considered in C , then there are cases $(x_0; a_0; r)$ in C which are distinguished only by the outcomes. In this case, it appears natural to assign the highest degree of similarity to these cases.

There are decision situations which are completely specified in the sense that all relevant aspects of the situation are included in the data x collected, as in Example 3.3. In such cases, one may be willing to assign similarity weights of zero to all cases with different data. This

appears to be an extreme case, which may obtain in experimental situations in statistics and physics. However, even in those applications, an insufficient number of relevant observations may necessitate considering data from similar, but not exactly identical situations. Such cases would naturally receive lower similarity weights $s(c \mid x_0; a_0)$.

In general, however, the decision maker will be uncertain about which data is important for the outcome. Such cases are described in Examples 3.1 and 3.2. In these cases, it would also be reasonable to consider cases in C which do not exactly match the decision situation $(x_0; a_0)$, but may be relevant given the lack of information about the data which determines the outcomes. In Example 3.2, there is data about a customer's choice of movie from a set of categories and languages. If the decision problem is to make a recommendation to a customer, it may be reasonable to give some weight to movies from the category and language which the customer has chosen in the past, but one may also want to consider other cases where customers with similar characteristics x bought movies from similar categories.

The following example will illustrate such a procedure for a variation of Example 3.1.

Example 3.1 resumed

Consider a medical doctor who has to choose one of two treatments, $a \in A = \{a_1; a_2\}$. In past treatments one has recorded only three characteristics of patients, high blood pressure, x_h , normal blood pressure, x_m , or low blood pressure, x_l . Hence, the set of potentially case-relevant data comprises also $x \in X = \{x_h; x_m; x_l\}$. Finally, three outcomes of the treatment have been registered, say r_1 , success, r_2 , no effect, and r_3 , failure, i.e., $R = \{r_1; r_2; r_3\}$.

In this case, databases D of any length T will be made up of the following eighteen cases:

$$C = \left\{ \begin{array}{ccc} \frac{c_1}{=} \frac{(x_l; a_1; r_1)}{=} & \frac{c_7}{=} \frac{(x_m; a_1; r_1)}{=} & \frac{c_{13}}{=} \frac{(x_h; a_1; r_1)}{=} \\ \frac{c_2}{=} \frac{(x_l; a_1; r_2)}{=} & \frac{c_8}{=} \frac{(x_m; a_1; r_2)}{=} & \frac{c_{14}}{=} \frac{(x_h; a_1; r_2)}{=} \\ \frac{c_3}{=} \frac{(x_l; a_1; r_3)}{=} & \frac{c_9}{=} \frac{(x_m; a_1; r_3)}{=} & \frac{c_{15}}{=} \frac{(x_h; a_1; r_3)}{=} \\ \frac{c_4}{=} \frac{(x_l; a_2; r_1)}{=} & \frac{c_{10}}{=} \frac{(x_m; a_2; r_1)}{=} & \frac{c_{16}}{=} \frac{(x_h; a_2; r_1)}{=} \\ \frac{c_5}{=} \frac{(x_l; a_2; r_2)}{=} & \frac{c_{11}}{=} \frac{(x_m; a_2; r_2)}{=} & \frac{c_{17}}{=} \frac{(x_h; a_2; r_2)}{=} \\ \frac{c_6}{=} \frac{(x_l; a_2; r_3)}{=} & \frac{c_{12}}{=} \frac{(x_m; a_2; r_3)}{=} & \frac{c_{18}}{=} \frac{(x_h; a_2; r_3)}{=} \end{array} \right\}$$

As in Example 3.1 and with a similar interpretation in mind, we can construct the sets of prob-

ability distributions $\hat{P}_T(c_i)$:

$$\begin{aligned}\hat{P}_T(c_i) &:= \{\hat{p} \in \Delta^2 \mid \hat{p}_1 \geq (1 - \frac{\varepsilon}{T})\} & \text{for } i = 1; 4; 7; 10; 13; 16 \\ \hat{P}_T(c_i) &:= \{\hat{p} \in \Delta^2 \mid \hat{p}_2 \geq (1 - \frac{\varepsilon}{T})\} & \text{for } i = 2; 5; 8; 11; 14; 17 \\ \hat{P}_T(c_i) &:= \{\hat{p} \in \Delta^2 \mid \hat{p}_3 \geq (1 - \frac{\varepsilon}{T})\} & \text{for } i = 3; 6; 9; 12; 15; 18\end{aligned}$$

for some $\varepsilon > 0$.

Assuming fixed similarity weights $(s_1; \dots; s_{18})$ for the eighteen basic cases, we arrive at the representation:

$$\begin{aligned}H_T(D) &: = \{h \in \Delta^2 \mid h = \left(\sum_{i=1}^{18} s(c_i) f_D(c_i) \right)^{-1} \left(\sum_{i=1}^{18} s(c_i) f_D(c_i) \hat{p}(c_i) \right), \\ & \quad i = 1; \dots; 18\}.\end{aligned}$$

Figure 2 illustrates this procedure. Disregarding ambiguity for the sake of the argument, let us assume for a moment that $\hat{P}_T(c = (x_i; a_j; r_k)) = \{e_k\}$ for all $c \in C$, where e_k denotes the k^{th} unit vector in the space of outcomes \mathbb{R}^3 .

Suppose that the problem under consideration is characterized by $(x_h; a_2)$. Then it appears natural to assign a similarity weight of one to the cases $\{c_{16}; c_{17}; c_{18}\}$ which are identical and differentiated only by the outcome. If one were to adhere strictly to the same type of cases, then one may want to put all other similarity weights to zero, leaving us with the relative frequencies of the outcomes observed for case $(x_h; a_2)$ in the sample D as the predicted probability over outcomes:

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = (f_D(c_{16}) + f_D(c_{17}) + f_D(c_{18}))^{-1} \begin{pmatrix} f_D(c_{16}) \\ f_D(c_{17}) \\ f_D(c_{18}) \end{pmatrix}.$$

For large data-sets D with many observations of cases $c_{16}; c_{17}; c_{18}$ this may be a reasonable procedure. It may well be, however, that D contains few or no observations of cases with high blood pressure, in which treatment a_2 has been prescribed. Then one may reasonably take into account cases which are not identical but arguably relevant. E.g., one could presume that a patient with normal blood pressure is more similar to a patient with high blood pressure than a patient with low blood pressure. Hence, the similarity of cases with normal blood pressure may be smaller than the similarity of cases with high blood pressure, but higher than the similarity of cases with low blood pressure, say one half. Assuming further that cases in which $x = x_l$ are

deemed irrelevant for this prediction yields:

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \left(\frac{1}{2}f_D(c_{10}) + \frac{1}{2}f_D(c_{11}) + \frac{1}{2}f_D(c_{12}) + f_D(c_{16}) + f_D(c_{17}) + f_D(c_{18}) \right)^{-1} \cdot \begin{pmatrix} f_D(c_{16}) + \frac{1}{2}f_D(c_{10}) \\ f_D(c_{17}) + \frac{1}{2}f_D(c_{11}) \\ f_D(c_{18}) + \frac{1}{2}f_D(c_{12}) \end{pmatrix}.$$

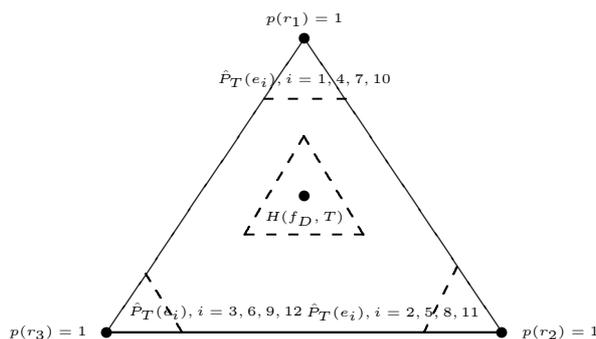


Figure 2. Constructing hypotheses using similarity weights

6 Learnability and confidence

The main focus of this paper is ambiguity in the context of case-based predictions of probabilities over outcomes. If a decision maker has to consider cases of different degrees of similarity then it appears natural to assume that a decision maker feels ambiguous about the predicted probability distribution over outcomes. There are several ways to model ambiguity, among them the multiple prior approach introduced by GILBOA & SCHMEIDLER (1989). In the spirit of this paper, we model ambiguity by a set of probability distributions over outcomes. The degree of ambiguity will be measured by set inclusion. The smaller the set of probability distributions over outcomes, the less ambiguous the prediction.

Notice that in Example 3.1, the sets $\hat{P}_T(c)$ shrink to a singleton if T tends to infinity, e.g.,

$$\lim_{T \rightarrow \infty} \hat{P}_T(x_i; a_j; r_k) = \{e_k\}$$

for all $x_i \in X$, $a_j \in A$, $r_k \in R$. Moreover, $\hat{P}_T(c_i) \subset \hat{P}_{T-1}(c_i)$ for all i . These assumptions

seem quite natural in the context of controlled experiments. The first one says that ambiguity decreases with "more information" in the sense of "more cases with the same outcome". The second one implies that as the same outcome is observed over and over again, its perceived probability converges to 1. In the following, we provide axioms which capture this intuition and analyze their implication for the perception of similarity.

We focus on a decision maker who tries to learn the properties of statistical experiments as those described in Example 3.3. In particular, we keep the characteristics of the situation as well as the action constant and model learning about the probability distribution of outcomes in R . 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 consists in formulating a set of probability distributions over outcomes, describing the likelihood of outcome r given a combination of an observed signal x and an action a . In the trivial case of a repeated experiment, i.e. $(x; a)$ is constant, the set of probability distributions over outcomes is assumed to contain the actually observed frequencies. The size of the set of probability distributions over outcomes can be taken to reflect the confidence of the decision maker with respect to the data. Given our assumption of ergodicity, as the data set becomes larger, the confidence of the decision maker increases until (with an infinite number of observations), 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 probability distribution $f(r)$ which exactly corresponds to the actual probability distribution of r given $(x; a)$:

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

This motivates the following axioms which specifies the rules by which the decision maker forms hypotheses.

Axiom A4 (*Learnability*) Consider databases with fixed $(x; a)$,

$$D = \left\{ (x; a; r_t)_{t=1}^T \right\}.$$

As $T \rightarrow \infty$,

$$H(D | x; a) \rightarrow \{h(D | x; a)\}$$

with

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

According to Axiom A4, 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.

Finally, we will assume that a decision maker's confidence in the observed frequencies of cases grows with a growing number of observations.

Axiom A5 (*Accumulation of knowledge*) Let $D = \{(x; a; r_t)_{t=1}^T\}$ and $D' = \{(x; a; r'_t)_{t=1}^{T'}\}$ be two finite data-sets such that $f_D = f_{D'}$ and $T' > T$, then

$$H(D' | x; a) \subset H(D | x; a).$$

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 in a controlled experiment, i.e., for fixed $(x; a)$. Notice that Axiom A5 applies only to data-sets in which *frequencies* are identical. If *frequencies* differ, a smaller set might be more reliable than a larger one. For example, $D \in \mathbb{D}_{100}$ with $f_D(x; a; r_1) = \frac{99}{100}$ and $f_D(x; a; r_2) = \frac{1}{100}$ will in general constitute stronger support for $h(r_1 | x; a) = \frac{99}{100}$ than $D' \in \mathbb{D}_{200}$ with $f_{D'}(x; a; r_1) = f_{D'}(x; a; r_2) = \frac{1}{2}$. Axiom A5 however does not tell us in which way the set of probabilities over outcomes shrinks. 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 6.1 *Assume Axioms A1, A4 and A5 hold, then for any database D of length T with fixed $(x; a)$, i.e., $D = ((x; a; r_t)_{t=1}^T)$, there is an $h \in H(D | x; a)$ such that*

$$h_r(D | x; a) = f_D(r)$$

for all $r \in R$.

Finally, we prove that together with the representation derived in Theorem 4.1, Axioms A4 and A5 imply two intuitive properties of the representation of $H(D)$. Firstly, the sets $\hat{P}_T(x; a; r)$ shrink with time, always contain the r -th unit vector e_r and converge to e_r as T converges to infinity. 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 6.2 *Suppose Axioms A4 and A5 hold. Then the representation $H(D | x; a)$ in Theorem 4.1 satisfies the additional properties:*

1. For all $r \in R$ and all T ,
 - (i) $\hat{P}_T((x; a; r) | x; a) \subset \hat{P}_{T-1}((x; a; r) | x; a)$,
 - (ii) $e_r \in \hat{P}_T((x; a; r) | x; a)$, and
 - (iii) $\lim_{T \rightarrow \infty} \hat{P}_T((x; a; r) | x; a) = \{e_r\}$.
2. For all $r \in R$, $\sum_{r \in R} s((x; a; r); (x; a)) = 1$.

7 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 data-bases 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 confirms 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 appears possible to find a representation of preferences over acts and a set of probabilities over outcomes depending on the data set.

Appendix A. Proofs

Proof of Theorem 4.1 :

It is obvious that the representation satisfies the axioms, hence we prove only 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 use of Proposition 3 from BGSS (2005, P. 1132), which we state in terms of our notation:

Proposition A.1 BGSS (2005) *Assume that $h : \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1} \rightarrow \Delta^{|R|-1}$ 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 consisting of finite data-bases in such a way that the limit of this sequence is a set of infinite data-bases \mathbb{D}_∞ . Moreover, we show that, for each vector f in the set $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$, \mathbb{D}_∞ contains a data-set, which has f as its frequency, 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 A.5), 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}_∞ . Next, in Lemma A.6, 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}_∞ . It is straightforward to show that the similarity weights do not depend on h . The last step, Lemma A.7, 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:

$$Q_T = \left\{ f \in \Delta^{|C|-1} \mid f(c) = \frac{k}{T} \text{ for some } k \in \{0; 1 \dots T\} \text{ and for all } c \in C \right\}.$$

Obviously, for each T , $Q_T \subset \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$. 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 $T_1; T_2 \dots T_m \dots$ with $T_m = m!$.

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

We will denote by \mathbb{D}_∞ the set of data-bases which give rise to the set $\mathbb{Q}_+^{|\mathcal{C}|} \cap \Delta^{|\mathcal{C}|-1}$.

Proof of Lemma A.2:

First, we show

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

Hence, we check that for each $q \in \mathbb{Q}_+^{|\mathcal{C}|} \cap \Delta^{|\mathcal{C}|-1}$, 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

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

with a_i and $b_i \in \mathbb{Z}^+$, and take the largest of the numbers b_i , $b(q) = \max_{i \in \{1 \dots |\mathcal{C}|\}} b_i$. Now set

$M = b(q)$ and observe that for all $m \geq M$, each ratio $\frac{a_i}{b_i}$ can be written as:

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

with

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

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:

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

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

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

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 $T \in \mathbb{Z}^+$, $f', f'', f \in Q_T$ 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 data-bases $D_1 = \dots = D_{m-1} = D_m = D'$; $D_{m+1} = \dots = D_n = D''$ 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:

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

and, hence, by (A2), there exists a vector $\mu \in \text{int}(\Delta^{n-1})$ 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. ■

For any $T \geq 2$, let H_T denote the restriction of H to \mathbb{D}_T . We now state a lemma which shows that for every such T , we can express

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

as a collection of single hypotheses (functions)

$$h_T : \mathbb{D}_T \rightarrow \Delta^{|R|-1},$$

$$h_T \in H_T$$

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

Lemma A.4 *Suppose that H_T satisfies (A2) and (A3). Then, for each $T \geq 2$, there is a set of functions*

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

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*

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

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

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

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 A.5 *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 $\mathbb{Q}_+^{|\mathcal{C}|} \cap \Delta^{|\mathcal{C}|-1}$ is replaced by Q_T for $T < \infty$.*

Proof of Corollary A.5:

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 $Q_\infty = \mathbb{Q}_+^{|\mathcal{C}|} \cap \Delta^{|\mathcal{C}|-1}$) and every rational $\alpha \in (0; 1)$, such that

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

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

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

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

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

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

Especially, for \mathbb{D}_∞ , this expression is valid for any two f and $f' \in \mathbb{Q}_+^{|\mathcal{C}|} \cap \Delta^{|\mathcal{C}|-1}$ 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 $\mathbb{R}^{|R|-1}$), each point of the set can be represented as a convex combination of at most $|R|$ points in $\mathbb{R}^{|R|-1}$. Since we have assumed that $H_T((c_i)^T)$ are convex sets (polyhedra), we can represent each such set as:

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

where $(\mu_{ij})_{j=1}^{|R|}$ is the above mentioned collection of points in $\mathbb{R}^{|R|-1}$. Note that since $((c_i)^T)_{i=1}^{|C|}$ 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 $(c_1)^T \dots (c_{|C|})^T$. By Lemma A.3, for every $D \in \mathbb{D}_T$,

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

with $\sum_{i=1}^{|C|} \lambda_i = 1$, $\lambda_i \in (0; 1)$, whenever c_i occurs in D at least once and $\lambda_i = 0$, else. 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 fixed collection of points $(\mu_{ij})_{i=1, j=1}^{|C|, |R|}$, we can identify each selection h_T with a vector of coefficients $(\alpha_{ij})_{i=1, j=1}^{|C|, |R|}$. 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 $(\alpha_{ij})_{i=1, j=1}^{|C|, |R|}$, the vectors:

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

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

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

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

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

for some $h_T\left((c_m)^T\right) \in H_T\left((c_m)^T\right)$. Whenever $h_T\left((c_i)^T\right)$, $h_T\left((c_j)^T\right)$ and $h_T\left((c_k)^T\right)$ entering this representation are non-collinear for any such $h_T(\hat{D})$, 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 represen-

tation are all collinear. If for each $m \in \{i; j; k\}$,

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

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

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

Now suppose that $h_T \left((c_m)^T \right)$ is an extreme point 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.

The last case to consider is the one in which $h_T \left((c_m)^T \right) \in \text{bd} \left(H_T \left((c_m)^T \right) \right)$, but are not extreme points for all $m \in \{i; j; k\}$. 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

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

and ϵ_j such that

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

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).

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 $h_T \left((c_m)^T \right)$ lie. It is obvious that ϵ_i and ϵ_j can always be chosen in such a way that

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

are not collinear.

If the three hyperplanes 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 $h_T \left((c_i)^T \right)$ belong to the interior of a face. If the edge containing, say $h_T \left((c_j)^T \right)$ is not collinear to the edge containing $h_T \left((c_k)^T \right)$, then, it is obviously possible to find ϵ_i and ϵ_j satisfying the necessary condition A-1. The idea is to move $h_T \left((c_j)^T \right)$ by ϵ_j along the edge to which it belongs, while moving the interior point $h_T \left((c_i)^T \right)$ 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 $h_T \left((c_j)^T \right)$ into the interior of $H_T \left((c_j)^T \right)$, whereas again it is always possible to move the interior point $h_T \left((c_i)^T \right)$ 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.6 *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

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

(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:

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

Proof of Lemma A.6:

Obviously, the construction in Lemma A.4 does not depend on T . Hence, for the sequence T_m as defined in Lemma A.2, letting $m \rightarrow \infty$, 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

$$\hat{p}_\infty(c_1) \dots \hat{p}_\infty(c_{|C|})$$

such that

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

or

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

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

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

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

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

separately by solving:

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

For the case $|C| = 3$, the condition that $h((c_1)^\infty)$, $h((c_2)^\infty)$ and $h((c_3)^\infty)$ 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 $\hat{p}_\infty(c_i) = h_\infty((c_i)^\infty)$ 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.7 For every $T \geq 2$ and $D \in \mathbb{D}_T$,

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

where

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

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

Proof of Lemma A.7:

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:

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

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 A.5 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:

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

and define

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

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

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

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

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

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 . And (A2) guarantees that for any $k \in \mathbb{Z}_+$,

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

implies

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

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

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

and from Lemma A.6, 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:

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

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}_∞ , for any \mathbb{D}_T with $T < \infty$. ■

Proof of Lemma 6.1:

Suppose that the frequency of r in a data-set $D = \{(x; a; r_t)_{t=1}^T\}$ is given by $f_D(r)$. Consider the sequence of data-sets D^k as $k \rightarrow \infty$ and note that by (A4), as $k \rightarrow \infty$,

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

By (A5), for each k ,

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

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

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

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

$$h_r(D) = f_D(r). \quad \blacksquare$$

Proof of Theorem 6.2:

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 $\hat{P}_T(c_i^{x;a} | x; a)$ by using only the data-set $(c_i^{x;a})^T$ and setting for each selection h ,

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

Hence,

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

(A5), *Accumulation of knowledge* ascertains that

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

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

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

The inclusion property shown above ascertains that

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

for every T . Now consider all cases $(x; a; r)_{r \in R}$ and the data sets $(x; a; r)^\infty$. For any two such sets with outcomes r_i and r_j , we know that

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

Now, expressing

$$H((x; a; r_i)^\infty \circ (x; a; r_j)^\infty | x; a)$$

in terms of similarity gives:

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

which implies

$$s((x; a); (x; a; r_i)) = s((x; a); (x; a; r_j)),$$

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

$$s((x; a); (x; a; r)) = 1$$

for all $r \in R$. ■

References

- 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.
- 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.
- 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. (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.
- KEYNES, J. M. (1921). *A Treatise on Probability*. London:Macmillan.
- PESKI, M. (2007). "Learning through Patterns", *mimeo*, University of Chicago.
- SAVAGE, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons.