ISSN 1936-5098

CAE Working Paper #07-03

Multiple Priors as Similarity Weighted Frequencies

by

Jürgen Eichberger
and
Ani Guerdjikova

April 2007

.

Multiple Priors as Similarity Weighted Frequencies

Jürgen Eichberger and Ani Guerdjikova

This version: January 22, 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. The set of probability distributions shrinks when new information confirms old data. 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 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 sets of data which may be more or less adequate for the decision problem under consideration. Hence, decision makers must aggregate data which may be more or less relevant for the unknown variable under consideration. The case-based decision-making approach of GILBOA & SCHMEIDLER (2001) offers a systematic approach to this information aggregation problem.

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 & 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 ambiguity about the probability distribution to be used for decision making. One may think of the set of probability distributions as those probability distributions the decision maker may not want to rule out, even given a most likely probability distribution. 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 resulting in one "black" and two "white" draws, the decision maker may be ambiguous about whether the probability is 0.5 for the two colors or whether there is a higher probability for 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.

Here we generalize the approach of BGSS (2005) such that it is possible to consider the weight of increasing evidence. With a growing number of observations, i.e., with the length of the database, decision makers may become more confident. Given a database, we will model ambiguity about the most likely probability distribution by a set of probability distributions rather

than a single probability distribution. We relax the main axiom of BGSS (2005), *Concatenation* to capture the idea that short data-sets provide ambiguous information about the actual probability distribution of outcomes. At the same time, our modification maintains the main property of the representation derived in BGSS (2005), the uniqueness of the similarity function. In a next step, we further specify the representation. In particular, we assume that the confidence of the decision-maker increases as data accumulates and that the set of probability distributions converges to the actual probability distribution when the data-set becomes very long.

As in BGSS (2005), it remains an open question which decision criterion one should use for choosing an action based on the available set of probability distributions over outcomes. The literature provides various decision criteria reflecting different degrees of optimism or pessimism in the face of ambiguity. Combining a decision rule with the information processing procedure described in this paper will be an issue of future research.

We view this paper as a first step in a broader research agenda. The topic of this paper is the relationship of ambiguity and similarity. In a second paper we will investigate the adjustment of the similarity function in the light of new information. The main issue will be the criterion in regard to which one wants to judge similarity. A third strand of research concerns the embedding of these ideas in a behavioral model in the spirit of GILBOA, SCHMEIDLER & WAKKER (2002).

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 data-sets, denoted by $\mathbb{D} := \underset{T \geq 1}{\cup} C^T$, 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 data base $D \in \mathbb{D}$. 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, it appears reasonable to assume that $H(D|x_0; a_0)$ is a compact and convex subset of $\Delta^{|R|-1}$. We will denote elements of this set by $h(D|x_0; a_0)$. For the probability assigned to outcome r by the probability distribution $h(D|x_0; a_0)$, we will write $h_r(D|x_0; a_0)$. 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 $H(D|x_0; a_0)$, $h(D|x_0; a_0)$, and $h_r(D|x_0; a_0)$, respectively.

2.1 Applications

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 2.1 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[1] *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, $H(D|x_0; a_0) \subset \Delta^{|R|-1}$, as possible. These probability distributions represent beliefs about the distribution 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 individually tailored recommendations. Our second example shows how an internet provider of a movie rental system can be modelled with this approach.

Example 2.2 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.,

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

[1] 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.

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. On the other hand, 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 is faced with drawings from an urn.

Example 2.3 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. Agents can bet on the color of the ball drawn, $\{B; W\}$*. Assume that players know the urn* x_0 *from which the ball is drawn, when they place their 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 players learn after each round of the lottery the result and *the urn from which the ball was drawn. Since there are only two bets possible* $a = 1_B 0$ *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, players have 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 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 $\mathrm{urn}\ 2$*, then he will evaluate the outcome of this bet based on the set of probability distributions $H(D|2; a_0)$.*

The set of probability distributions $H(D|2; a_0)$ should reflect both the decision maker's information given by the database 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 $\mathrm{urn}\ 2$ *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 may will depend on the number of relevant observations $T(2) = f_D(2, B) + f_D(2, W)$, e.g.,*

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

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 as less relevant for a draw from urn 2[2].

In the following sections, we will take the decision situation (x_0, a_0) as given. We will relate

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

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 $H(D|x_0, a_0)$. We will impose axioms on the set of probability distributions over outcomes $H(D|x_0; a_0)$ which will imply a representation of the following type,

$$H(D|x_0; a_0) = \left\{ \frac{\sum_{c \in D} s(c|x_0, a_0) f_D(c) h_c}{\sum_{c \in D} s(c|x_0, a_0) f_D(c)} \mid h_c \in H((c)^{|D|} | x_0, a_0) \right\},$$

where $(c)^{|D|}$ denotes a database of length $|D|$ containing only case c.

The weighting function $s(c|x_0, a_0)$ represents the perceived similarity between the case c and the current situation (x_0, a_0). It indicates how relevant a case c is with respect to the decision situation (x_0, a_0). The set of probability distributions over outcomes $H((c)^{|D|} | x_0, a_0)$ is the set of probability distributions over outcomes entertained by the decision maker in case of a data set consisting only of observations of the same case c. The axioms suggested below will imply (up to a normalization) unique similarity weights $s(c|x_0, a_0)$ and unique sets of probability distributions $H((c)^{|D|} | x_0, a_0)$. This result generalizes the main theorem of BGSS (2005) to the case of multiple priors.

It appears natural to assume that a decision maker with a database consisting only of observations of the same case $c = (x, a, r)$ will at least consider the possibility that the outcome r occurs with probability 1 in the decision situation (x, a), i.e., that the r-th unit vector $e_r \in \mathbb{R}^{|R|}$ is contained in the set $H((x_0, a_0, r)^{|D|} | x_0, a_0)$. Moreover, confirming evidence should increase the confidence in this belief,

$$\lim_{|D| \to \infty} H((x_0, a_0, r)^{|D|} | x_0, a_0) = \{e_r\}.$$

We will also provide axioms for these properties of the set of probability distributions $H(D|x_0; a_0)$.

3 Axioms and Representation

In this section, we will take the decision situation (x_0, a_0) as given and will 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, which will be 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 $\pi : \{1...T\} \to \{1...T\}$, 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 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. 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 3.1 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 $D^k = \underbrace{D \circ ... \circ D}_{k\text{-times}}$ 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 $H\left((c)^{10000}\right)$ 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 contains also probability distributions $h((c))$ with $h_{r'}((c)) \in (0, 1)$ for all r'. In particular, based on the database $D = (c)$, 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 $H\left((c)^{10000}\right)$.

We would like to model decision problems where decision makers may become more confident about their beliefs as they observe databases with the same frequency distribution of cases but increasing numbers of cases. Hence, 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 $D_1 ... D_n \in \mathbb{D}_T$ be such that $D_1 \circ ... \circ D_n = F^n$. Then, there exists a vector $(\lambda_1 ... \lambda_{n-1}) \in int(\Delta^{n-1})$ such that, for every $k \in \mathbb{Z}^+$,

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

To understand 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 $D_2 = D_3 = (e_2)^3$:

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

$(A2)$ then asserts the existence of λ_1 and λ_2 such that:

$$\lambda_1 H(D_1) + \lambda_2 H(D_2) + (1 - \lambda_1 - \lambda_2) H(D_3) = H(F).$$

It should be clear that each data-set 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 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 $\mathbb{D}_T$, $(A2)$ has the following implication: if the evidence of each of the n data-sets of equal length $D_1 ... D_n$ suggests that a given outcome r is possible, i.e. for all $i \in \{1...n\}$,

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

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 $\lambda \in (0, 1)$

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

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

Remark 3.1 *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 $\mathbb{D}_T$, $(c_1)^T, ..., \left(c_{|C|}\right)^T$ 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:

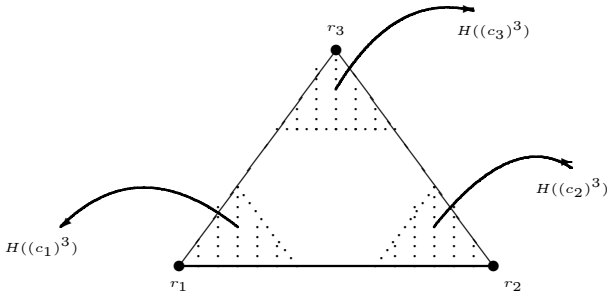
– either singletons

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

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 $(A4)$ 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 3.1 *Let H be a correspondence $H : \mathbb{D} \to \Delta^{|R|-1}$ the images of which are non-empty convex and compact sets. Then the following two statements are equivalent:*

(i) *H satisfies the Axioms* Invariance, Concatenation, *and* Linear Independence *for every T.*

(ii) *There exists a function*

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

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

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

such that for any $D \in \mathbb{D}_T$

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

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 $T < \infty$.

The following example of a correspondence H will illustrate the result.

3.1 Leading example

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 case, databases D of any length $|D|$ will be made up of the following twelve cases:

c_1	$=$	(x_1, a_1, r_1)	c_7	$=$	(x_2, a_1, r_1)
c_2	$=$	(x_1, a_1, r_2)	c_8	$=$	(x_2, a_1, r_2)
c_3	$=$	(x_1, a_1, r_3)	c_9	$=$	(x_2, a_1, r_3)
c_4	$=$	(x_1, a_2, r_1)	c_{10}	$=$	(x_2, a_2, r_1)
c_5	$=$	(x_1, a_2, r_2)	c_{11}	$=$	(x_2, a_2, r_2)
c_6	$=$	(x_1, a_2, r_3)	c_{12}	$=$	(x_2, a_2, r_3)

Recall that, without any loss of generality, we can replace any database$D$ with $(f_D, |D|)$. Hence, one can write (e_i, T) for a database $D = (c_i)^T$, which contains T-times the case c_i. For each T and each unit vector $e_i \in \mathbb{R}^{|C|-1} = \mathbb{R}^{11}$, consider the following sets of probabilities over outcomes,

$$\begin{aligned}
\hat{P}_T(e_i) &:= \{p \in \Delta^2 |\, p_1 \geq (1 - \tfrac{\varepsilon}{T})\} \quad \text{for} \quad i = 1, 4, 7, 10, \\
\hat{P}_T(e_i) &:= \{p \in \Delta^2 |\, p_2 \geq (1 - \tfrac{\varepsilon}{T})\} \quad \text{for} \quad i = 2, 5, 8, 11, \\
\hat{P}_T(e_i) &:= \{p \in \Delta^2 |\, p_3 \geq (1 - \tfrac{\varepsilon}{T})\} \quad \text{for} \quad i = 3, 6, 9, 12
\end{aligned}$$

for some $\varepsilon > 0$.

This assignment of probabilities over outcomes can be given the following interpretation. An intuitive way would be to assign probability distributions to databases would be as follows. 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 decrease as the number of confirming observations T rises. 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 $\hat{P}_T(e_i)$ satisfy Linear independence (Axiom $(A3)$).

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

$$H(f_D, T) := \{p \in \Delta^2 |\, p = \left(\sum_{i=1}^{12} s_i f_D(c_i)\right)^{-1} \left(\sum_{i=1}^{12} s_i f_D(c_i) h_i\right),\ h_i \in \hat{P}_T(e_i),\ i = 1, ..., 12\}.$$

4 Similarity

The similarity weights $s(c_i)$ of Theorem 3.1 have to be seen in relation to the decision situation $(x_0; a_0)$ under consideration. The notation $s\,(c|x_0, a_0)$ 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 2.3. In such cases, one may be willing to assign similarity weights of zero to all cases with different data. This appears as an extreme case, which may obtain in experimental situations in statistics and physics. Even in those applications, there may be insufficient observations. A lack of the desired data may make it sensible to consider data from similar, but not exactly equal situations. Hence, one may want to include cases with data from similar situations with lower similarity weights $s\,(c|x_0, a_0)$.

In general, however, decision makers will be uncertain about which data will be important for the outcome. Such cases are described in the Examples 2.1 and 2.2. In these cases, it may be reasonable to consider also 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 2.2, 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 maybe from the same language group bought other categories.

The following example will illustrate such a procedure for a variation of Example 2.1.

Example 4.1 *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 $|D|$ will be made up of the following eighteen cases:*

c_1	$=$	(x_l, a_1, r_1)	c_7	$=$	(x_m, a_1, r_1)	c_{13}	$=$	(x_h, a_1, r_1)
c_2	$=$	(x_l, a_1, r_2)	c_8	$=$	(x_m, a_1, r_2)	c_{14}	$=$	(x_h, a_1, r_2)
c_3	$=$	(x_l, a_1, r_3)	c_9	$=$	(x_m, a_1, r_3)	c_{15}	$=$	(x_h, a_1, r_3)
c_4	$=$	(x_l, a_2, r_1)	c_{10}	$=$	(x_m, a_2, r_1)	c_{16}	$=$	(x_h, a_2, r_1)
c_5	$=$	(x_l, a_2, r_2)	c_{11}	$=$	(x_m, a_2, r_2)	c_{17}	$=$	(x_h, a_2, r_2)
c_6	$=$	(x_l, a_2, r_3)	c_{12}	$=$	(x_m, a_2, r_3)	c_{18}	$=$	(x_h, a_2, r_3)

Analogously to Example 3.1 (and with a similar interpretation in mind), we can construct the sets of probability distributions $\hat{P}_T(e_i)$ in the following way.

$$\hat{P}_T(e_i) := \{p \in \Delta^2 |\, p_1 \geq (1 - \tfrac{\varepsilon}{T})\} \quad \textit{for} \quad i = 1, 4, 7, 10, 13, 16$$
$$\hat{P}_T(e_i) := \{p \in \Delta^2 |\, p_2 \geq (1 - \tfrac{\varepsilon}{T})\} \quad \textit{for} \quad i = 2, 5, 8, 11, 14, 17$$
$$\hat{P}_T(e_i) := \{p \in \Delta^2 |\, p_3 \geq (1 - \tfrac{\varepsilon}{T})\} \quad \textit{for} \quad i = 3, 6, 9, 12, 15, 18$$

for some $\varepsilon > 0$.
Assuming fixed similarity weights $(s_1, ..., s_{18})$ for the eighteen basic cases, we arrive at the representation:

$$H(f_D, T) := \{p \in \Delta^2 |\, p = \left(\sum_{i=1}^{18} s_i f_D(c_i)\right)^{-1} \left(\sum_{i=1}^{18} s_i f_D(c_i) h_i\right),\ h_i \in \hat{P}_T(e_i),\ i = 1, ..., 18\}.$$

Figure 2 illustrates this procedure. Disregarding ambiguity for the sake of the argument, let us assume for a moment that $\hat{P}_T(e_i) = \{e_i\}$ for $i = 1, ..., 18$.
Similarity weights have to be interpreted in relation to a given situation (x_0, a_0) of which the probability over results has to be assessed. If there are cases in C which are characterized by the same (x_0, a_0) a similarity relation among the basic cases in C can be established.
E.g., 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

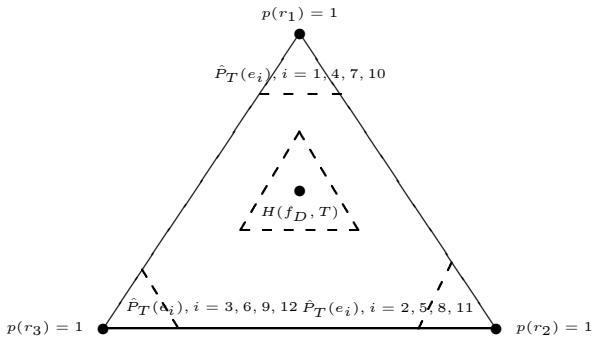
of the outcomes observed for case (x_h, a_2) in the sample D as the predicted probability over outcomes:

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

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 test cases 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. This would yield

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

■



2. Constructing hypotheses using similarity weights

In this paper we assume that similarity weights for a given reference case are independent of the amount of data in D. This assumption appears, however, questionable if one views the perception of similarity as an imperfect substitute for knowledge about the relevance of underlying data. This is an issue we will deal with in a companion paper, EICHBERGER & GUERDJIKOVA (2007).

The main focus of thes paper is ambiguity in the context of case-based predictions of probabil-

ities 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 the sets $\hat{P}_T(e_i)$ shrink to a singleton if T tends to infinity, e.g.,

$$\lim_{T\to\infty} \hat{P}_T(e_i) = \{e_i\}$$

for all i. Moreover, $\hat{P}_T(e_i) \subset \hat{P}_{T-1}(e_i)$ for all i.These assumption seems 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 next section, we provide axioms which capture this intuition and analyse their implication for the perception of similarity.

5 Learnability and confidence

In this section, we focus on a decision-maker who tries to learn the properties of statistical experiments as in Example 2.3 in the previous section. 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):

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

Of course, it is easy to think of examples in which the ergodicity property would not be satisfied. E.g. the sequence of observations $(x; a; 1); (x; a; 2) ... (x; a; 100) ... (x; a; 200) ...$ does not have the ergodicity property. Learning from this sequence would have a completely different character than the one incorporated in our axioms.

This motivates the assumption of *Learnability* which we make below. We assume that the following axioms specify 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 $T \to \infty$,

$$H(D) \to \{h(D)\}$$

with

$$h_r(D) = f_D(r).$$

According to Axiom $(A4)$, the decision maker can learn the unknown proportion of the colors in a urn, as in Example 2.3. If draws from the urn are with replacement, then the decision-maker will eventually learn the true composition of the urn after observing the outcome of an infinite number of draws.

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 and D' be two finite data-sets with common (x, a) such that $f_D = f_{D'}$ and $|D'| > |D|$, then

$$H(D') \subset H(D).$$

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 experi-

ment, 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 $f_{D'}(x; a; r_1) = f_{D'}(x; a; r_2) = \frac{1}{2}$. Note that $(A5)$ does not tell us in which way the set of probabilities over outcomes shrinks. Together with the *Invariance Axiom* (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 $(A1)$, $(A4)$ 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 $h \in H(D)$ such that*

$$h_r(D) = f_D(r)$$

for all $r \in R$.

Finally, we prove that together with the representation derived in Theorem 3.1, Axioms $(A4)$ and $(A5)$ imply two intuitive properties of the representation of $H(D)$. First, 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. Second, 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 and consider databases D with fixed (x, a), then the representation $H(D)$ in Theorem 3.1 satisfies the following additional properties:*

1. *$\hat{P}_T$ satisfies for all $r \in R$ and every T,*
   (i) $\hat{P}_T((x, a, r)) \subset \hat{P}_{T-1}((x, a, r))$,
   (ii) $e_r \in \hat{P}_T((x, a; r))$, *and*
   (iii) $\lim_{T\to\infty} \hat{P}_T((x; a; r)) = \{e_r\}$.
2. $s((x, a; r)) = 1$ *for every $r \in R$.*

6 Concluding remarks

We have generalized the approach of BGSS (2005) to understand the influence of ambiguity on a decision maker's prediction about the probability distribution of outcomes. 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, holds as long as we impose some consistency on the weights λ across different values of T. This consistency is essential for the uniqueness result. Relaxing this assumption would require the similarity function to be determined separately for each value of T, i.e. for each set $\mathbb{D}_T$. As a result, for $T < \infty$, different similarity functions would be used to evaluate different data-sets D. Along the sequence $\mathbb{D}_{T_m}$ with $T_m = m!$, the set of similarity functions will shrink, approaching a single point in the limit as $m \to \infty$.

As a special case of our approach we consider predictions associated with homogenous data-sets. Homogenous data-sets can be interpreted as controlled statistical experiments, hence the idea that ambiguity decreases as new data confirms past evidence appears very natural. Combined with the assumption that in the limit, the decision-maker learns the probability distribution generating the process, we arrive at the conclusion that the notion of similarity in such situations is trivial. In particular, all observations are considered equally important for the prediction to be made.

Although statistical experiments can serve as an illustration of our approach, we do not consider them to be the ideal field for the application of the concept of similarity. Rather, we think that our three examples (2.1, 2.2, 2.3) provide an illustration of the type of situations, in which similarity is useful in the process of reasoning about probabilities. In our current framework, it might be natural to use the structure of the sets X and A to construct more specific similarity functions. However, such a construction is reasonable only if the structure imposed on the sets reflects the structure of payoffs, in other words, $(x; a)$ pairs which are considered similar lead to

"similar" probability distributions over outcomes. It is easy to imagine situations, in which the decision-maker first has to *find out* which of the characteristics of $(x; a)$ are payoff-relevant (e.g. for a child, drawing on a piece of paper and drawing on the wall might appear quite similar, until it learns that it earns compliments for the former, but reproach for the latter). Hence, the data-set will provide not only information about the distribution of payoffs of a specific alternative, but also about similarity between alternatives. The more observations it contains, the more precise the perception of similarity will become. We plan to model this adaptive process in a subsequent paper.

Similarly to BGSS (2005), our approach relies on concepts which are not directly observable, such as probability or similarity perceptions. Using the axiomatizations of the case-based decision rule, as well as of decision-making rules under ambiguity present in the literature, it is possible to extend the results of this paper so as to accommodate actual choices. We plan to address this issue in our future research.

Appendix A. Proofs

Proof of Theorem 3.1 :

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 make use of the following Proposition 3 from BGSS (2005, P. 1132), which we state in terms of our notation:

Proposition 6.1 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 posi-*

tive 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 $\mathbb{D}_\infty$. Moreover, we show that, for each vector f in the set $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$, $\mathbb{D}_\infty$ contains a data-set, which has f as its frequency, see Lemma 6.2. 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 6.3, 6.4 and Corollary 6.5), 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 6.1 when restricted to $\mathbb{D}_\infty$. Next, in Lemma 6.6, 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 $\mathbb{D}_\infty$. It is straightforward to show that the similarity weights do not depend on h. The last step, Lemma 6.7, 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:

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

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 6.2 *Consider the infinite sequence* $T_1; T_2...T_m...$ *with* $T_m = m!$.

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

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

Proof of Lemma 6.2:

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 6.3 allows us to relate the *Concatenation Axiom*, $(A2)$ (which is formulated in terms of data-sets) to property (i) in Proposition 6.1 (stated in terms of frequencies).

Lemma 6.3 *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 6.3:

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

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

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

as a collection of single hypotheses (functions)

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

which satisfy properties (i) and (ii) of Proposition 6.1.

Lemma 6.4 *Suppose that H_T satisfies $(A2)$ and $(A3)$. Then, for each $T \geq 2$, there is a set of 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*

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

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 6.4, we illustrate its implications by the following corollary:

Corollary 6.5 *Each $h_T \in \mathcal{H}_T$ as constructed in Lemma 6.4 satisfies properties (i) and (ii) stated in Proposition 6.1, where the set $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ is replaced by Q_T for $T < \infty$.*

Proof of Corollary 6.5:

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 6.1. To see the relation between (i') and (i) recall that Lemma 6.3 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

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

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

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 6.4:

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 $H_T\left((c_i)^T\right)$ are convex sets (polyhedra),

we can represent each such set as:

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

where $\left(\mu_{ij}\right)_{j=1}^{|R|}$ is the above mentioned collection of points in $\mathbb{R}^{|R|-1}$. Note that since $\left(\left(c_i\right)^T\right)_{i=1}^{|C|}$ 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 $\left(c_1\right)^T...\left(c_{|C|}\right)^T$. By Lemma 6.3, for every $D\in\mathbb{D}_T$,

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

with $\lambda_i\in(0;1)$, whenever c_i occurs in D at least once. The Caratheodory Theorem now allows us to write any such convex combination as:

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

Hence, 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:

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

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 $\left(h_T^{\hat{D}}\left(D\right)\right)_{D\in\mathbb{D}_T}$ at last three of which are non-collinear.

Suppose first that H_T satisfies the condition of $(A3)$ for some $\left(c_i\right)^T$, $\left(c_j\right)^T$ and $\left(c_k\right)^T$, all of

which are single points:

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

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

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

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 $H_T\left(\left(c_m\right)^T\right)$ for $m \in \{i; j; k\}$ have a non-empty interior. Take some set

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

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

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

for some $h_T\left(\left(c_m\right)^T\right) \in H_T\left(\left(c_m\right)^T\right)$. Whenever $h_T\left(\left(c_i\right)^T\right)$, $h_T\left(\left(c_j\right)^T\right)$ and $h_T\left(\left(c_k\right)^T\right)$ entering this representation are non-collinear for any such $h_T\left(\hat{D}\right)$, the result of the lemma obtains. Suppose, however that $h_T\left(\left(c_i\right)^T\right)$, $h_T\left(\left(c_j\right)^T\right)$ and $h_T\left(\left(c_k\right)^T\right)$ entering the representation are all collinear. If for each $m \in \{i, j, k\}$,

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

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

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

and

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

Now suppose that $h_T\left(\left(c_m\right)^T\right)$ is an extreme point of $\left(H_T\left(\left(c_m\right)^T\right)\right)$ 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 $h_T\left((c_m)^T\right) \in bd\left(H_T\left((c_m)^T\right)\right)$, 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

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

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).

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 $h_T\left(\left(c_i\right)^T\right)$ 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 6.6 *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

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

(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 6.6:

Obviously, the construction in Lemma 6.4 does not depend on T. Hence, for the sequence T_m as defined in Lemma 6.2, letting $m \to \infty$, we can represent H_∞ as a selection of functions h_∞ which satisfy all of the conditions of Proposition 6.1. 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

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

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 6.4 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 is does not depend of the chosen vector and is given by:

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

Lemma 6.7 *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

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

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

Proof of Lemma 6.7:

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:

$$\frac{\frac{T-1}{T}s\left(c_1\right)\hat{p}_\infty\left(c_1\right)+\frac{1}{T}s\left(c_2\right)\hat{p}_\infty\left(c_2\right)}{\frac{T-1}{T}s\left(c_1\right)+\frac{1}{T}s\left(c_2\right)} \tag{A-3}$$

$$= \lambda^1 h_\infty\left(\left(c_1\right)^\infty\right)+\left(1-\lambda^1\right)h_\infty\left(\left(c_2\right)^\infty\right) \tag{A-4}$$

...

$$\frac{\frac{T-1}{T}s\left(c_{|C|-1}\right)\hat{p}_\infty\left(c_{|C|-1}\right)+\frac{1}{T}s\left(c_{|C|}\right)\hat{p}_\infty\left(c_{|C|}\right)}{\frac{T-1}{T}s\left(c_{|C|-1}\right)+\frac{1}{T}s\left(c_{|C|}\right)}$$

$$= \lambda^{|C|-1}h_\infty\left(\left(c_{|C|-1}\right)^\infty\right)+\left(1-\lambda_1^{|C|-1}\right)h_\infty\left(\left(c_{|C|}\right)^\infty\right) \tag{A-5}$$

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 6.5 shows that properties (i) and (ii) stated in Proposition 6.1 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:

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

and define

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

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

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

and conclude, by $(A2)$ and Lemma 6.3 that

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

for some $\lambda\in(0;1)$. Lemma 6.4 shows that the same values of λ can be used for each selection h_T of H_T. And $(A2)$ guarantees that for any $k\in\mathbb{Z}_+$,

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

implies

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

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

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

and from Lemma 6.6, we know that:

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

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:

$$\frac{\frac{T-1}{T}s\left(c_1\right)\hat{p}_T\left(c_1\right)+\frac{1}{T}s\left(c_2\right)\hat{p}_T\left(c_2\right)}{\frac{T-1}{T}s\left(c_1\right)+\frac{1}{T}s\left(c_2\right)} \tag{A-6}$$

$$= \lambda^1 h\left(\left(c_1\right)^T\right)+\left(1-\lambda^1\right)h\left(\left(c_2\right)^T\right) \tag{A-7}$$

...

$$\frac{\frac{T-1}{T}s\left(c_{|C|-1}\right)\hat{p}_T\left(c_{|C|-1}\right)+\frac{1}{T}s\left(c_{|C|}\right)\hat{p}_T\left(c_{|C|}\right)}{\frac{T-1}{T}s\left(c_{|C|-1}\right)+\frac{1}{T}s\left(c_{|C|}\right)}$$

$$= \lambda^{|C|-1}h\left(\left(c_{|C|-1}\right)^T\right)+\left(1-\lambda^{|C|-1}\right)\left(h\left(c_{|C|}\right)^T\right) \tag{A-8}$$

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 6.1 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 $\hat{p}_T\left(c\right)$ as long as

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

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 $D=\left\{\left(x;a;r_t\right)_{t=1}^T\right\}$ is given by $f_D\left(r\right)$. Consider

the sequence of data-sets D^k as $k \to \infty$ and note that by $(A1)$, 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 see that the proposition holds note that we construct the elements of $\hat{P}_T\left(\left(\bar{x};\bar{a}\right);c_i\right)$ by using only the data-set $\left(\left(c_i\right)^T\right)$ and setting for each selection h,

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

Hence,

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

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

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

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

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

The inclusion property shown above ascertains that

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

for every T. Now consider all cases $\left(x^i;a^i;r\right)_{r\in R}$ and the data sets $\left(x^i;a^i;r\right)^{\infty}$ which assign a

frequency one to $(x^i; a^i; r)$. For any two such sets, we know that

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

Now, expressing

$$H\left(\left(x^i;a^i;r\right)^\infty \circ \left(x^i;a^i;r'\right)^\infty\right)\left(x^i;a^i\right)$$

in terms of similarity gives:

$$\begin{aligned} &\left(0;0...\underbrace{\frac{1}{2}}_{r^{\text{th}}\text{ position}};0...\underbrace{\frac{1}{2}}_{r'^{\text{th}}\text{ position}};0...0\right) \\ = &\frac{\frac{1}{2}s\left(\left(x_i;a_i\right);\left(x_i;a_i;r'\right)\right)e_r + \frac{1}{2}s\left(\left(x_i;a_i\right);\left(x_i;a_i;r\right)\right)e_r}{\frac{1}{2}s\left(\left(x_i;a_i\right);\left(x_i;a_i;r'\right)\right) + \frac{1}{2}s\left(\left(x_i;a_i\right);\left(x_i;a_i;r\right)\right)}, \end{aligned}$$

which implies

$$s\left(\left(x_i;a_i\right);\left(x_i;a_i;r\right)\right) = s\left(\left(x_i;a_i\right);\left(x_i;a_i;r'\right)\right),$$

for all r, $r' \in R$, which after normalization can be written as:

$$s\left(\left(x_i;a_i\right);\left(x_i;a_i;r\right)\right) = 1$$

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.

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.), North-Holland Publishing Company, Amsterdam.

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

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