

Eichberger, Jürgen; Guerdjikova, Ani

**Conference Paper**

## Case-Based Belief Formation under Ambiguity

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session:  
Asymmetric Information and Incentives, No. A8-V3

**Provided in Cooperation with:**

Verein für Socialpolitik / German Economic Association

*Suggested Citation:* Eichberger, Jürgen; Guerdjikova, Ani (2010) : Case-Based Belief Formation under Ambiguity, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2010: Ökonomie der Familie - Session: Asymmetric Information and Incentives, No. A8-V3, Verein für Socialpolitik, Frankfurt a. M.

This Version is available at:

<https://hdl.handle.net/10419/37467>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Case-Based Belief Formation under Ambiguity

Jürgen Eichberger  
University of Heidelberg

and

Ani Guerdjikova  
Cornell University

23 October 2009

This version: October 23, 2009

## Abstract

In this paper, we consider a decision-maker who tries to learn the distribution of outcomes from previously observed cases. For each observed database of cases the decision-maker predicts a set of priors expressing his beliefs about the underlying probability distribution. We impose a version of the concatenation axiom introduced in BILLOT, GILBOA, SAMET, AND SCHMEIDLER (2005) which ensures 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. The predicted probabilities, however, may vary with the number of observations. This generalisation of BILLOT, GILBOA, SAMET, AND SCHMEIDLER (2005) allows one to model learning processes.

**JEL Classification:** D81, D83

**Keywords:** case-based decision theory, ambiguity, multiple priors, learning, similarity

**Address for Correspondence:** Ani Guerdjikova, Cornell University, Department of Economics, 462 Uris Hall, [ag334@cornell.edu](mailto:ag334@cornell.edu).

---

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

# 1 Introduction

How will existing information influence probabilistic beliefs? How do data enter the inductive process of determining a prior probability distribution? KEYNES (1921) 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.

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

In a recent paper, BILLOT, GILBOA, SAMET, AND SCHMEIDLER (2005), henceforth BGSS (2005), show that, under few assumptions, a probability distribution over outcomes can be derived as a similarity-weighted average of the frequencies of observed cases. Moreover, GILBOA, LIEBERMAN, AND SCHMEIDLER (2006) 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. Furthermore, the probability distribution depends only on the frequency of observations in the data, but not on the length of the database. This approach appears satisfactory 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 particular, this approach restricts the decision maker to being a frequentist, but allows the weights assigned to the frequencies to depend on the perceived relevance of the cases. Two important aspects of the decision situation are, however, neglected. First, even if the decision maker is able and willing to assign a probability distribution to each database, this distribution might vary both with the frequency and *the length* of the database, as for instance in Bayesian updating. Second, in the face of ambiguity, the decision maker might find himself unable to pinpoint a unique probability distribution.

In this paper, we modify the approach of BGSS (2005) in two ways. First, we allow the prediction of the decision maker to depend both on the frequency and on the length of the database. This allows us to capture the idea that, controlling for the frequency, longer databases contain more precise information and to incorporate Bayesian updating as a special case of our analysis. Second, we allow the predictions to be represented by a convex set of probability distributions to capture the idea that information can be ambiguous.

O'HAGAN AND LUCE (2003) describe the difficulty of making and interpreting point predictions about probabilities as follows:

"The first difficulty we will face is that the expert will almost certainly not be an expert in probability and statistics. That means it will not be easy for this person to express her beliefs in the kind of probabilistic form demanded by Bayes' theorem. Our expert may be willing to give us an estimate of the parameter, but how do we interpret this? Should we treat it as the mean (or expectation) of the prior distribution, or as the median of the distribution, or its mode, or something else? [...] We could go on to elicit from the expert some more features of her distribution, such as some measure of spread to indicate her general level of uncertainty about the true value of the parameter." (pp. 64-65).

While decision makers might be unable to make point predictions about a prior distribution, they may be able to identify a range of possible probabilities, either directly as upper and lower bounds of probabilities, or indirectly by a degree of confidence expressed regarding a point prediction. In the former case, a convex set of probabilities is suggested directly, in the latter

case, one may view the set of probabilities as a neighborhood of an imprecise point prediction. The decision maker's ambiguity can be related to the length of the database (insufficient number of observations) or to the content of the data (observations which do not exactly correspond to the case for which a prediction has to be made). The first type of ambiguity is relevant even in the perfect case of randomized statistical experiments. Consider, e.g., a decision maker observing random draws with replacement from an urn containing black and white balls in unknown proportions. After one white and two black balls have been drawn out of the urn, the decision maker might entertain a set of priors describing his beliefs about the constitution of the urn. This set might include the observed frequency  $(\frac{1}{3}; \frac{2}{3})$ , but it might also contain other distributions, e.g.,  $(\frac{1}{2}; \frac{1}{2})$ , if the decision maker considers the number of observations to be insufficient to generate an exact prediction. This type of ambiguity will disappear as the number of observations grows.

The second type of ambiguity arises in situations in which the data contains relevant, but not identical cases to the one, for which a prediction has to be made. A case often discussed in the econometrics literature is the one of missing variables, see MANSKI (2000), as well as Example 1 in GONZALES AND JAFFRAY (1998). For instance, medical studies often contain large sets of data, but fail to record potentially important characteristics, such as the gender of the patients. MANSKI (2000) argues that in this case, the probability distribution over outcomes cannot be point-identified. Hence, the data is consistent with a set of probability distributions which will be non-degenerate even as the number of observations becomes large.

Our model provides a representation which allows for both types of ambiguity. To obtain this, we modify the main axiom of BGSS (2005), *Concatenation*, by restricting it to databases of equal length, i.e., thus controlling for the ambiguity resulting from insufficient amount of data. In order to establish a connection between the predictions for databases of different lengths, we introduce an additional axiom — "Learning". It captures the idea that, as information accumulates, the changes in the forecast caused by additional confirming observations get smaller and smaller. If the perceived ambiguity is due to a limited number of observations, the set of pre-

dictions converges to a singleton. However, the axiom does not exclude the case of persistent ambiguity, when predictions converge to a non-degenerate set of beliefs. This distinction between ambiguity which vanishes with a sufficiently large number of observations and ambiguity which remains for any number of observations corresponds to a similar distinction in EPSTEIN AND SCHNEIDER (2007).

Despite these modifications, our approach obtains a similarity function which is unique and independent of the content and the size of the databases. This property is a central feature of the representation in BGSS (2005). Hence, as in BGSS (2005), frequentism and kernel classification represent special cases of our representation. However, our approach captures a broader scope of rules, including Bayesianism and full Bayesian updating on a set of priors. Moreover, it allows us to model persistent ambiguity arising from missing or inadequate data. There exist several approaches as to how a decision maker, whose forecast consists of a set of probability distributions, can select a prior from this set<sup>1</sup>. Among these, the max-min rule suggested by GILBOA AND SCHMEIDLER (1989) and MANSKI (2000) is by far the most popular. It selects the probability distribution which results in minimal expected utility for the specific action. Our approach allows us to establish a connection between the set of priors and the selection of a single probability distribution used to evaluate a specific act. We show that whenever the sets of priors which the decision maker associates with different databases satisfy the axioms of our model, so do the probability distributions determined according to the max-min rule.

As in BGSS (2005), the question remains open which decision criterion one should use given the decision maker's beliefs. In order to obtain a decision rule together with a multiple prior representation one may embed these ideas in a behavioral model in the spirit of GILBOA, SCHMEIDLER, AND WAKKER (2002) or derive decision criteria reflecting degrees of optimism or pessimism in the face of ambiguity as in the work of COIGNARD AND JAFFRAY (1994) and GONZALES AND JAFFRAY (1998). We propose such a behavioral approach in EICHBERGER

---

<sup>1</sup> For instance, a Bayesian might assign a prior on the set of possible probability distributions and take expectations with respect to this prior. Alternatively, one can use the center of the set of probability distributions (the Steiner point) as a focal probability distribution, see GAJDOS, HAYASHI, TALLON, AND VERGNAUD (2007).

AND GUERDJIKOVA (2008). We believe, however, that a characterization of the mapping  $H$  from databases to probabilities over outcomes is desirable in its own right. It opens up the possibility to study the optimal use of data for the derivation of a set of prior distributions and to model databased learning rules.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 outlines the model and Section 4 provides some motivating examples. In Section 5, we state the axioms. Section 6 presents the main result. In Section 7, we collect some examples which illustrate our approach and show that it is compatible with an array of popular statistical methods. Section 8 concludes the paper. All proofs are collected in the Appendix.

## 2 Related Literature

There are several ways to model ambiguity of a decision maker in the literature. A representation of ambiguous beliefs by means of capacities was introduced by SCHMEIDLER (1989). For convex capacities, this approach coincides with the multiple prior approach advanced in GILBOA AND SCHMEIDLER (1989). BEWLEY (1986) derives a set of probability distributions from incomplete preferences. These multiple-prior approaches were developed further by GHIRARDATO, MACCHERONI, AND MARINACCI (2004) and Chateauneuf, Eichberger, and Grant (2007). KLIBANOFF, MARINACCI, AND MUKERJI (2005) model ambiguity attitudes by a second-order probability distribution over a set of probability distributions. All these multiple-prior approaches represent ambiguity by a set of probability distributions which a decision maker considers when evaluating her expected utility. In the spirit of these models, we model ambiguity by a set of probability distributions over outcomes. The degree of ambiguity can be measured by set inclusion. The smaller the set of probability distributions over outcomes, the less ambiguous the prediction.

In GILBOA AND SCHMEIDLER (1989) the set of priors is purely subjective. In contrast, several recent papers, AHN (2008), GAJDOS, HAYASHI, TALLON, AND VERGNAUD (2007), STINCHCOMBE (2003), provide a framework to analyze decisions in situations in which the

set of priors is objectively given. This allows them to distinguish objectively given Knightian uncertainty from the subjective attitude towards ambiguity. In our framework, a similar distinction is achieved differently. The decision maker associates with each database a set of probability distributions, which take into account both the objective information contained in the data (i.e. the nature and frequency of cases observed, as well as the number of observations) and the subjective degree of ambiguity. Thus, our approach provides a method to characterize sets of subjective priors related to the data-generating process.

MARINACCI (2002) and EPSTEIN AND SCHNEIDER (2007) analyze statistical learning in the context of ambiguity. MARINACCI (2002)'s work provides conditions under which ambiguity almost surely "fades away" as data accumulate. In contrast, EPSTEIN AND SCHNEIDER (2007) distinguish two types of scenarios: one in which it is possible to learn the objective probability distribution and another where ambiguity is persistent. They study the effect of prior information on the learning process in the context of statistical experiments in the spirit of ELLSBERG (1961). If information about the colors of a given number of balls in an urn is obtained from "sampling with replacement", such "learning" will reveal the proportions of colors in the long run. In contrast, if the composition of the color in the urn is changing over time, e.g., in Scenario 3 of EPSTEIN AND SCHNEIDER (2007) (p. 1279) because a certain number of balls is replaced by an administrator in every period, then learning by sampling with replacement cannot reveal the true proportions of colors and ambiguity will prevail even in the long run. EPSTEIN AND SCHNEIDER (2007) show also how these types of ambiguity induce different investment behavior in a portfolio choice model.

Our framework uses the notion of similarity to distinguish between controlled statistical experiments, and situations in which relevant, but not completely identical cases have been observed. If the observed cases are identical to the case under consideration, as in a controlled statistical experiment, e.g., in Scenarios 1 and 2 of EPSTEIN AND SCHNEIDER (2007), then the decision-maker will be able to learn the objective probability distribution satisfying the ergodicity property. When, however, the observed cases are distinct from the situation under consideration, as

in Scenario 3, then ambiguity may persist in form of a limit set of probabilities, even if a large number of data has been collected. Yet, sampling may still provide information, though the decision maker has to judge its relevance based on some presumption about the administrator's behavior.

GONZALES AND JAFFRAY (1998) model preferences over Savage-type acts for a given set of, possibly imprecise, data. They derive a representation of preferences in form of a linear combination of the maximal and the minimal potential outcome of an act and its expected utility with respect to the observed frequency of states. The weights attached to the maximal and minimal outcomes can be interpreted as degrees of optimism and pessimism. They decrease over time relative to the weight attached to the expected utility part of the representation. Because observations may be imprecise a decision maker associates with a set of data a set of priors centered around the observed frequency. The size of the set of probabilities depends negatively on the amount of data. While we do not derive a decision rule from behavior, our approach encompasses a richer class of situations which is not restricted to the case of controlled statistical experiments considered in both COIGNARD AND JAFFRAY (1994) and GONZALES AND JAFFRAY (1998).

### 3 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 *databases of length*  $T$  is denoted by  $\mathbb{D}^T := C^T$ . Finally,  $\mathbb{D} := \bigcup_{T \geq 1} \mathbb{D}^T$  denotes the set of databases of arbitrary length.

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

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

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

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

We interpret  $H(D \mid x_0; a_0)$  as the set of probability distributions over outcomes which the decision maker takes into consideration given the database  $D$ .

## 4 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 4.1 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, gender, age, medical history, etc. Having observed the characteristics  $x_0$ , the physician evaluates a treatment  $a_0$  based on the assessment of the probability distribution over outcomes  $r \in R$ . A set of cases  $D$  observed<sup>2</sup> in the past may serve the physician in this assessment of probabilities over outcomes.*

---

<sup>2</sup> 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.

A case  $c = (x_t; a_t; r_t)$  is a combination of a patient  $t$ 's characteristics  $x_t$ , treatment assigned  $a_t$  and outcome realization  $r_t$  recorded in the database  $D$ . Given the database  $D$ , the physician considers a set of probabilities over outcomes,  $H(D | x_0; a_0) \subset \Delta^{|R|-1}$ , as possible. These probability distributions represent beliefs about the likelihood of possible outcomes after choosing a treatment  $a_0$  for the patient with characteristics  $x_0$ .

In general, the physician will form his beliefs based on cases in which characteristics potentially different from  $x_0$  and actions potentially different from  $a_0$  were observed. E.g., O'HAGAN AND LUCE (2003)(pp. 62-64) discuss how information from different studies about the effectiveness of similar, but not identical, drugs can be combined into a prior distribution.

Two problems can prevent the physician from specifying a unique probability distribution for a specific treatment. First, he might have few observations, and, therefore, doubt that the observed frequencies are representative of the population as a whole. Second, the observations might not be identical to the case at hand (e.g., the physician might have a vast amount of data on patients with flu symptoms, which allows him to evaluate different treatments, however, he might consider all these cases to be of only limited relevance when faced with the symptoms of swine flu). While in the first situation, collecting more data of the same type would reduce the ambiguity, in the second, the ambiguity is due to incomplete understanding of the relation between different cases.

Consider, e.g. a situation in which some of the characteristics contained in  $X$  were not recorded. As in MANSKI (2000), suppose that in a study that contains the outcomes of a specific treatment, the gender of the patient is not recorded<sup>3</sup>. Suppose that the treatment resulted in success for exactly 50% of all cases. A physician who has to assign a treatment to a woman will not be able to infer from the database which of the following scenarios corresponds to the truth: (i) the treatment is always effective for men, but never for women; (ii) the treatment is always effective for women, but never for men; (iii) the treatment is successful in 50% of cases, for both genders (of course many more intermediate cases are possible). Even after observing a very large database, the physician will be completely ignorant of the probability of success, and his prediction will be represented by the interval  $[0; 1]$ . Here, the fact that the cases in the data

---

<sup>3</sup> If patient's characteristics are represented by a vector  $x = (x^1 \dots x^I)$ , we could use a default value of  $\bar{x}^i$  to denote that characteristic  $i$  has not been recorded for the specific observation.

are not completely identical to the case at hand gives rise to sets of probability distributions. ■

As a second application we will consider classic statistical experiments where the decision maker bets on the color of the ball drawn from an urn.

#### **Example 4.2** Lotteries

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

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

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

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

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

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

Note that for a given urn  $x$ , the observation of a case, allows the decision maker to observe the outcome of the actually chosen action, but also to infer the (counterfactual) outcome of the lottery he did not choose. This is a specific feature of this example, which distinguishes it from Example 4.1.

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 payoffs  $\{1; 0\}$  given a bet  $a$ . Suppose a decision maker with the information of database  $D$  learns that a ball will be drawn from urn 2

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

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

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

Example 4.2 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<sup>4</sup> for a draw from urn 2. The representation derived in the next section allows for this possibility.

## 5 Axioms

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

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

---

<sup>4</sup> Part III of KEYNES (1921) provides an extensive review of the literature on induction from cases to probabilities.

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 set of probability distributions over outcomes  $\hat{P}_T^{(c \mid x_0; a_0)}$  denotes the beliefs of the decision maker when the database  $D = \underbrace{(c \dots c)}_{T\text{-times}}$  is observed. This set may depend on the number of observations. It may be large for small numbers and may shrink as more confirming data become available. The weighting function  $s(c \mid x_0; a_0)$  represents the relevance of a case  $c$  for the current situation  $(x_0; a_0)$  and can be interpreted as the perceived similarity between  $c$  and  $(x_0; a_0)$ .

The axioms we introduce below imply that  $s(\cdot \mid x_0; a_0)$  is unique (up to a normalization) and does not depend on  $T$ , while the sets of probability distributions  $\hat{P}_T^{(c \mid x_0; a_0)}$  are determined uniquely. This result generalizes the main theorem of BGSS (2005) to the case in which beliefs depend on the number of observations and can be expressed as sets of probability distributions over outcomes.

In the following discussion,  $(x_0; a_0)$  is assumed constant. Hence, we suppress notational reference to it and write  $H(D)$ ,  $h(D)$ ,  $\hat{P}_T^c$  and  $s(c)$  instead of  $H(D \mid x_0; a_0)$ ,  $h(D \mid x_0; a_0)$ ,  $\hat{P}_T^{(c \mid x_0; a_0)}$  and  $s(c \mid x_0; a_0)$ , respectively. It is important to keep in mind, however, that all statements of axioms and conclusions do depend on the relevant reference situation  $(x_0; a_0)$ . In particular, the similarity weights, deduced below, measure similarity of cases relative to this reference situation.

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

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

**Remark 5.1** By Axiom (A1), we can identify every database  $D = (c_t)_{t=1}^T$  with the corresponding multi-set<sup>5</sup>  $\{(c_t)_{t=1}^T\}$ , in which the number of appearances of every case  $c$  exactly corresponds to the number of its appearances in  $D$ . We will denote the database and its corresponding multi-set by the same letter. In particular, when we write  $D = D'$ , we mean equality of the multi-sets corresponding to the databases  $D$  and  $D'$ .

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

**Definition 5.1** (Concatenation) For any  $T, T' \in \mathbb{N}$ , and any two databases  $D = (c_t)_{t=1}^T$  and  $D' = (c'_t)_{t=1}^{T'}$ , the database

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

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

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

**Notation**  $D^k = \underbrace{D \circ \dots \circ D}_{k\text{-times}}$  denotes  $k$  concatenations of the same database  $D$ . In particular, a database consisting of  $k$ -times the same case  $c$  can be written as  $c^k$ .

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  $D, D' \in \mathbb{D}$ ,

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

---

<sup>5</sup> On multi-sets see, e.g., BLIZARD (1988).

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

This axiom can be easily adapted to our framework:

**Axiom BGSS – Multiple Priors** (*Concatenation with multiple priors*) For every  $D, D' \in \mathbb{D}$

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

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

Both versions of the axiom imply that, for any  $k$ , the databases  $D$  and  $D^k$  map into the same set of probability distributions over outcomes,  $H(D) = H(D^k)$ . Hence, two databases  $D = c$  and  $D' = c^{10000}$  will be regarded as equivalent. This seems counter-intuitive. 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. Suppose, that we restrict the prediction to be single-valued, e.g., because the decision maker is a Bayesian. Unless, the decision maker's prior assigns a probability of 1 to outcome  $r$ , this decision maker will assign a higher probability to outcome  $r$  under  $D'$  than under  $D$ . If, in contrast, the decision maker considers the situation to be ambiguous, we could argue that the database  $c^{10000}$  provides a strong evidence for a probability distribution concentrated on the outcome  $r$ ;  $h_r(c^{10000}) = 1$ . Based on a single observation  $(x; a; r)$ , however, it appears quite reasonable to consider a set of probability distributions  $H(c)$  which also contains probability distributions  $h(c)$  with  $h_{r'}(c) \in (0,1)$  for all  $r'$ . In particular, based on the information contained in  $D = (c)$ , a decision maker may not be willing to exclude the case of all outcomes being equally probable, i.e.,  $\bar{h}(D)$  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 concatenations of arbitrary databases  $D$  and  $D'$ . Restricting the axiom to databases with equal length will provide sufficient flexibility for our purpose.

To illustrate the idea, consider two cases  $c_1$  and  $c_2$  and databases with two observations of these cases, say  $D_1 = (c_1, c_1)$ ,  $D_2 = (c_2, c_2)$ , and  $F = (c_1, c_2)$ . Due to Axiom (A1), one

can write these databases in terms of frequencies and numbers of observations as  $F = (f_F, 2)$ ,  $D_1 = (f_{D_1}, 2)$ , and  $D_2 = (f_{D_2}, 2)$ . Since  $D_1 \circ D_2 = (c_1, c_1, c_2, c_2) = F \circ F$  holds, the frequency of cases in  $F$  must be a mixture of the frequencies of  $D_1$  and  $D_2$ ,

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

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

Axiom (A2) generalizes this idea: for any  $n$  databases of equal length  $T$  that can be concatenated to an  $n$ -fold of a database  $F$  of length  $T$ , we postulate that any probability distribution over outcomes predicted on the basis of database  $F$  can be expressed as a convex combination of probability distributions over outcomes associated with the databases  $D_i$ .

**Axiom A2** (*Concatenation restricted to databases of equal length*) Consider databases  $F \in \mathbb{D}^T$  and  $D_1 \dots D_n \in \mathbb{D}^T$  for some  $n \in \mathbb{N}$ , such that  $D_1 \circ \dots \circ D_n = F^n$ . Then, there exists a vector  $\lambda \in \text{int}(\Delta^{n-1})$  such that

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

In spirit, Axiom (A2) is very similar to the Concatenation Axiom introduced by BGSS (2005). It has the following intuitive interpretation<sup>6</sup>: if a decision maker cannot exclude a certain probability distribution  $h$  after observing the evidence in any of the databases  $D_1 \dots D_n$ , then he should not be able to exclude it after observing the evidence in a database of the same length,  $F$ , the frequency of which is a mixture of the frequencies of  $D_1 \dots D_n$ . The main difference to the Concatenation Axiom of BGSS (2005) is that we restrict the axiom to databases of equal length.

---

<sup>6</sup> Note that the Axiom *does not have* the following behavioral implication: if action  $a$  is preferred to  $a'$  under all databases  $D_1 \dots D_n$ , then it is also preferred under  $F$ . To understand this, consider the case of  $n = 2$ . Let  $a \succsim_{D_1} a'$  and  $a \succsim_{D_2} a'$ . Suppose also that the evidence contained in database  $D_1$  is more relevant for  $a$ , while the evidence contained in  $D_2$  is more relevant for  $a'$ . Suppose that, at the same time, the decision maker values  $a'$  higher given the relevant evidence contained in  $D_2$  than he values  $a$ , given the relevant evidence for this action,  $D_1$ . In this case, combining the evidence contained in the two databases  $D_1$  and  $D_2$  into  $F$  might lead to a reversal of preferences, i.e.,  $a' \succsim_F a$ . The same argument applies also for the Concatenation Axiom of BGSS (2005).

The restriction to sets of equal length is important for our approach since databases with identical frequencies, but different length may give rise to different sets of probabilities over outcomes. In particular, depending on some learning rule (e.g. full Bayesian updating, see Section 7), it may be reasonable to assume that the set of probabilities is non-degenerate, but converges towards the observed frequency of outcomes as more observations of the same cases become available. Intuitively, the decision maker becomes more confident that the observed frequencies reflect the actual data-generating process for the database  $D^{k+1}$  than for  $D^k$ . In contrast, applying the *Concatenation Axiom* of BGSS (2005), we would have to conclude that for some  $\lambda \in \text{int}(\Delta^k)$ ,

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

for all  $k \in \mathbb{N}$ . Thus, imposing BGSS (2005)'s *Concatenation Axiom*, the set of probability distributions over outcomes would necessarily be independent of the number of observations. Our weaker Axiom (A2), however, implies in this case only  $\sum_{i=1}^{k+1} \lambda_i H(D) = H(D)$ , which is trivially satisfied for any set  $D$ .

Axiom (A2) allows us to identify the similarity function. In general, however, similarity will depend on the length of the database. In order to prevent this, we impose

**Axiom A3** (*Constant Similarity*) Consider the databases  $F \in \mathbb{D}^T$  and  $D_1 \dots D_n \in \mathbb{D}^T$  for some  $n \in \mathbb{N}$ , such that  $D_1 \circ \dots \circ D_n = F^n$ . If for some  $\lambda \in \text{int}(\Delta^{n-1})$ ,

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

holds for some  $k \in \mathbb{N}$ , then it holds for any  $k \in \mathbb{N}$ .

Independence of the similarity function from the number of observations is justified if one assumes that the similarity of cases is determined by some primitive knowledge about the cases, which is not based on the information contained in the database<sup>7</sup>. We discuss this axiom and its

---

<sup>7</sup> Compare also the discussion of "structural priors" in O'HAGAN AND LUCE (2003) (pp. 67-68).

implications in Section 7.4.

The following axiom requires learning processes to be stable. If the number of observations of the same case  $c$  case increases, beliefs about the outcome of  $(x_0; a_0)$  will react less to each additional observation and will eventually settle on a (possibly singleton) set of probability distributions.

**Axiom A4** (*Learning*) For every  $c \in C$ , the sequence of sets  $H_T(c^T)$  converges<sup>8</sup>.

We will use the notation  $H_\infty(c)$  for the limit of the sequence  $\lim_{T \rightarrow \infty} H_T(c^T)$ . Since all sets  $H_T(c^T)$  are non-empty, compact and convex subsets of  $\Delta^{|R|-1}$ , the limit  $H_\infty(c)$  inherits these properties.

Under Axiom (A4), ambiguity may persist or vanish in the limit depending on the similarity of cases to the situation under consideration. E.g., if  $c$  represents a statistical experiment w.r.t. the action of interest, i.e.,  $c = (x_0; a_0; r)$  for some  $r \in R$ , then it appears reasonable to assume that  $H_\infty(c) = \{\delta_r\}$ , where  $\delta_r$  is the Dirichlet distribution putting mass 1 on  $r$ . However, if  $c$  includes the observation of an action distinct from  $a_0$ , say,  $a'$ , there is no reason to suggest that the decision maker will be able to eliminate all ambiguity about the performance of action  $a$  even after observing an infinite sequence of realizations of  $a'$ . Hence, in general, the limit set will contain more than one element.

The next axiom requires those elements of  $(H_\infty(c))_{c \in C}$  which are singletons or segments to be non-collinear.

**Axiom A5** (*Non-collinearity*) No three of the sets  $H_\infty(c)$  of dimension 0 or 1 are collinear.

Axiom (A5) replaces the Axiom Non-collinearity in BGSS (2005). While BGSS (2005) require that there are at least three non-collinear vectors in the set  $(h(D))_{D \in \mathbb{D}}$ , our restriction is imposed on the limit sets  $H_\infty(c)$ . To understand the restrictions imposed by (A5), it is useful to first look at its implications in the setting of BGSS (2005). If  $h$  is a function and the Concate-

<sup>8</sup> For the definition of set convergence, see Definition 4.1 in ROCKAFELLAR AND WETS (2004). Since the sets  $H(c^T)$  are subsets of the  $|R| - 1$ -dimensional simplex, it follows that in our model, this notion of convergence coincides with convergence with respect to the Pompeiu-Hausdorff distance, see Example 4.13 in ROCKAFELLAR AND WETS (2004).

nation Axiom of BGSS is satisfied, we know that  $h_\infty(c)$  exists and  $h_\infty(c) = h(c^T) = h(c)$  for all  $c \in C$ . This means that no three of the predictions related to the basic cases are collinear. Intuitively, this excludes the possibility that the set of basic cases  $C$  can be reduced by taking the evidence from a given case  $c$  to be exactly equivalent to the evidence of a database containing observations of two different cases,  $c'$  and  $c''$  in a certain proportion. This requirement is satisfied, for controlled randomized experiments, i.e., for the cases of the type  $(x_0; a_0; r)$ , for which ambiguity vanishes and the limit prediction can be reasonably assumed to be  $\delta_r$ .

However, ambiguity need not vanish for cases in which characteristics distinct from  $x_0$  and  $a_0$  have been observed. In this case, the only restriction imposed by Axiom (A5) concerns those sets  $H_\infty(c)$  which are segments. We require that they are not collinear to any other two segments or points in the set  $(H_\infty(c))_{c \in C}$ . No assumptions are imposed on those sets in  $(H_\infty(c))_{c \in C}$  with a dimension 2 or higher.

## 6 Representation Theorem

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

**Theorem 6.1** *Let  $H$  be a correspondence  $H : \mathbb{D} \rightarrow \Delta^{|R|-1}$  the images of which are non-empty convex, and compact sets and which satisfies the Axioms (A4) Learning and (A5) Non-collinearity. 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 (A1) Invariance, (A2) Concatenation restricted to databases of equal length, and (A3) Constant Similarity.*

(ii) *There exists a unique, up to multiplication by a positive number, function*

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

*and a unique correspondence*

$$\hat{P} : \{2, 3, \dots\} \times C \rightarrow \Delta^{|R|-1}$$

such that for all  $T \geq 2$  and any  $D \in \mathbb{D}^T$ ,

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

The proof of the Theorem is relegated to the Appendix. Note, however, how the axioms affect the representation. Axiom (A1) ensures that the prediction associated with a database depends only on the number and the frequency of the observations, but not on the order in which they arrive. Axiom (A2) implies the existence of weights  $s_T(c)$  such that the predicted probability distributions associated with a database  $D \in \mathbb{D}^T$  can be expressed as a weighted average of the predictions of the individual cases in this set. Without imposing further restrictions, the weights  $s_T(c)$  will depend on the length of the database  $T$  and be non-unique. Axiom (A3) yields independence of the similarity weights of the number of observations. Axioms (A4) and (A5) ensure uniqueness of the representation.

As in BGSS (2005), uniqueness cannot be guaranteed unless some of the predictions corresponding to cases in  $C$  are non-collinear. In contrast to the framework of BGSS (2005), in which the predictions from databases consisting of observations of a single case  $c$  are independent of the number of observations  $T$ , here the predictions depend on the number of observations. Hence, in order to deduce a unique similarity function, one could require non-collinearity for every value of  $T$ . An alternative, and in our opinion more intuitive approach, which is chosen here, is to ensure that the predictions from databases of increasing length converge to some limit set, Axiom (A4), and to guarantee non-collinearity of limit sets which are singletons or segments, Axiom (A5). Then, for sufficiently large databases, there exists a selection of at least three predictions which are not collinear and which allow us to identify the similarity function uniquely.

Note that replacing Axioms (A2) and (A3) by the Axiom BGSS–Multiple Priors in Theorem 6.1 would imply that the correspondence  $\hat{P}$ , and hence the predictions  $H(D)$ , would be independent of the length of the database.

For some applications, the forecast of a decision maker about the outcome of a given action may be a unique probability distribution. Using the lower-case letter  $h$  to indicate the special case

where  $H$  is a function rather than a correspondence, it is straightforward to rewrite our axioms for this special case. In particular, Axiom (A4), Learning, now reduces to the requirement that for each  $c \in C$ , the sequence  $h_T(c^T)$  converges to some probability distribution  $h_\infty(c)$ . Non-collinearity, Axiom (A5), now ensures that no three of these limit vectors will be collinear. Hence, we obtain the following corollary to Theorem 6.1:

**Corollary 6.2** *Let  $h$  be a function  $h : \mathbb{D} \rightarrow \Delta^{|R|-1}$  which satisfies Axioms (A4) and (A5) 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 restricted to databases with equal length (A2) and Constant Similarity (A3).*

(ii) *There exists a unique, up to multiplication by a positive number, function*

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

*and a unique function*

$$\hat{p}_T : \{2, 3, \dots\} \times C \rightarrow \Delta^{|R|-1}$$

*such that for all  $T \geq 2$  and any  $D \in \mathbb{D}^T$ ,*

$$h_T(D) = \frac{\sum_{c \in C} s(c) f_D(c) \hat{p}_T^c}{\sum_{c \in C} s(c) f_D(c)}.$$

Allowing the predicted probability distribution to depend on the length of a database, Corollary 6.2 generalizes the result of BGSS (2005). The time-dependency of this representation allows us to model learning processes. For example, with increasing numbers of observations the predicted probability distribution may become less sensitive to new additional data. The following section provides examples from statistical models.

## 7 Examples and Applications

A special case of our approach are predictions based on homogenous databases which contain the same characteristics and actions in all observations. Hence, all data have the same similarity. Homogenous databases result typically from controlled statistical experiments. In this context, it appears natural to assume that ambiguity decreases as new data confirm past evidence.

In this section we show by examples that several statistical procedures satisfy the axioms of Theorem 6.1. Moreover, we discuss situations where constant similarity appears natural and illustrate how our method can be used to select a probability distribution from a set of priors.

## 7.1 Frequentism

Consider a decision maker who observes the outcome of a statistical experiment, where the set of possible cases is given by  $C = ((x_0; a_0; r))_{r \in R}$ . After observing a database  $D$  of length  $T$  and frequency  $f$ , the decision maker's beliefs about the outcome of action  $a_0$  are described by  $h_T(D) = f$ . It is easy to check that this rule satisfies all the axioms. This prediction rule has the special property that the prediction does not depend on the length of the database, but only on the observed frequency of cases.

Should the set of cases include also pairs of characteristics and actions different from  $(x_0; a_0)$  then the decision maker predicts a probability distribution  $\hat{p}^{(x;a;r)}$  over outcomes of action  $a_0$  in circumstances  $x_0$  the for each observed case  $(x; a; r)$ . The similarity weight  $s(x; a; r)$  describes the relevance of this case  $(x, a; r)$  for the prediction about  $(x_0; a_0)$ . This is the case axiomatized by BGSS (2005).

## 7.2 Bayesianism and Full Bayesian Updating

Bayesian updating is one of the most prominent statistical learning rules. Its generalization to full Bayesian updating incorporates learning with multiple priors, see MARINACCI (2002). In both cases, predictions depend on the observed frequency as well as on the length of the database. Hence, neither of these rules satisfies the Concatenation Axiom formulated by BGSS (2005). Here, we show that both Bayesianism and full Bayesian updating constitute special cases of our approach.

As in MARINACCI (2002), consider a decision maker who is trying to learn the probability distribution over the outcomes in a statistical experiment where sampling takes place with replacement.

The set of possible cases is given by  $C = ((x_0; a_0; r))_{r \in R}$ . Let  $D$  be a database of length  $T$ . Then  $Tf_D(r)$  is the number of observations of  $r$  in  $D$ . Suppose that the decision maker's

prior information is reflected by an initial set of priors  $\mathcal{P}$ , consisting of Dirichlet distributions on  $\Delta^{|R|-1}$ . Then  $\mathcal{P}$  can be described by the (strictly positive) parameters of these distributions,  $(\alpha_1, \dots, \alpha_{|R|})$ . In particular, for a Dirichlet distribution with parameter  $\alpha_k$ , the expected probability of outcome  $r$  in absence of any observations is given by  $\frac{\alpha_r}{\sum_{k=1}^{|R|} \alpha_k}$ . The initial set of distributions  $H_0$  is given by:

$$\left\{ \left( \frac{\alpha_1}{\sum_{k=1}^{|R|} \alpha_k}, \dots, \frac{\alpha_{|R|}}{\sum_{k=1}^{|R|} \alpha_k} \right) \mid (\alpha_1 \dots \alpha_{|R|}) \in \mathcal{P} \right\}.$$

The Bayesian update of a Dirichlet probability distribution on  $\Delta^{|R|-1}$  with parameters  $(\alpha_1 \dots \alpha_{|R|})$  after observing a database  $D$  is another Dirichlet distribution with parameters

$$(\alpha_1 + T f_D(r_1), \dots, \alpha_{|R|} + T f_D(r_{|R|})).$$

Hence, full Bayesian updating on the set  $\mathcal{P}$  implies that the decision maker updates each of the priors according to the Bayesian rule,

$$H(D) = \left\{ \left( \frac{\alpha_1 + T f_D(r_1)}{\sum_{k=1}^{|R|} \alpha_k + T}, \dots, \frac{\alpha_{|R|} + T f_D(r_{|R|})}{\sum_{k=1}^{|R|} \alpha_k + T} \right) \mid (\alpha_1, \dots, \alpha_{|R|}) \in \mathcal{P} \right\}.$$

Standard Bayesian updating obtains as a special case where  $\mathcal{P}$  is a singleton and  $H(D)$  is the probability distribution obtained by Bayesian updating.

Note that the order in which information arrives does not affect the posterior, hence Axiom (A1) is satisfied. Furthermore, let  $F, D_1 \dots D_n$  be databases of length  $T$  such that

$$F^n = D_1 \circ \dots \circ D_n.$$

There are strictly positive coefficients  $(\gamma_i)_{i=1}^n$  summing to 1 such that

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

These  $(\gamma_i)_{i=1}^n$  are independent of  $T$ . Hence, we have

$$\begin{aligned} H(F) &= \left\{ \left( \frac{\alpha_1 + T f_F(r_1)}{\sum_{k=1}^{|R|} \alpha_k + T}, \dots, \frac{\alpha_{|R|} + T f_F(r_{|R|})}{\sum_{k=1}^{|R|} \alpha_k + T} \right) \mid (\alpha_1, \dots, \alpha_{|R|}) \in H_0 \right\} \\ &= \left\{ \left( \frac{\alpha_1 + T \sum_{i=1}^n \gamma_i f_{D_i}(r_1)}{\sum_{k=1}^{|R|} \alpha_k + T}, \dots, \frac{\alpha_{|R|} + T \sum_{i=1}^n \gamma_i f_{D_i}(r_{|R|})}{\sum_{k=1}^{|R|} \alpha_k + T} \right) \mid (\alpha_1, \dots, \alpha_{|R|}) \in H_0 \right\} \\ &= \sum_{i=1}^n \gamma_i \left\{ \left( \frac{\alpha_1 + T f_{D_i}(r_1)}{\sum_{k=1}^{|R|} \alpha_k + T}, \dots, \frac{\alpha_{|R|} + T f_{D_i}(r_{|R|})}{\sum_{k=1}^{|R|} \alpha_k + T} \right) \mid (\alpha_1, \dots, \alpha_{|R|}) \in H_0 \right\} = \sum_{i=1}^n \lambda_i H(D_i), \end{aligned}$$

with  $\lambda_i =: \gamma_i$  for  $i \in \{1 \dots n\}$ . Full Bayesian updating, therefore, satisfies Axiom (A2). Since

the coefficients  $(\gamma_i)_{i=1}^n$ , and hence also  $(\lambda_i)_{i=1}^n$ , do not depend on  $T$ , Axiom (A3) is satisfied as well.

Since the parameters  $\alpha_1 \dots \alpha_n$  are strictly positive, it follows that all priors contained in  $H_0$  assign strictly positive probabilities to all outcomes in  $R$ . Therefore, each of the sequences  $H \left( (x_0; a_0; r)^T \right)$  will converge to the unit vector assigning probability 1 to outcome  $r$ . Hence, Axiom (A4), is satisfied as well. Finally, since the limit sets coincide with the unit vectors in  $\Delta^{|R|-1}$ , Axiom (A5) holds.

### 7.3 Kernel Density Classification

In contrast to the previous examples, kernel methods are not restricted to databases generated from the same statistical experiment. Therefore, similarity plays a role. Consider the standard kernel density classification model<sup>9</sup>.

There are  $|R|$  classes and a set of objects each of which can be described by a vector of characteristics  $x \in X$ . For instance, a physician might want to divide his patients into classes according to their reaction to a certain type of drug<sup>10</sup>. The physician observes cases of the form  $c = (x; r)$ , in which a patient of type  $x$  has been classified into class  $r$ . The physician entertains a notion of closeness between the patients described by a similarity function  $s : C \rightarrow \mathbb{R}$ . Suppose that the reaction of a patient of type  $x_0$  has to be classified. The information of the physician is given by a database  $D$  of length  $T$ .

The kernel density classification proceeds as follows<sup>11</sup>. Given the database  $D$  which contains the cases in which patients of different types have been classified, one needs to determine the relevance of these cases to the case at hand, i.e., the similarity of an observed patient  $x$  to  $x_0$ . Weighting the frequencies of cases in which outcome  $r$  has been observed by their similarities

<sup>9</sup> See, e.g., HASTIE, TIBSHIRANI, AND FRIEDMAN (2001)(p. 184).

<sup>10</sup> The classification may also be "action-dependent", e.g., one might be interested in classifying the patients relative to their risk of contracting a specific disease, conditional on a treatment they have undergone (e.g. vaccination).

<sup>11</sup> A similar problem showing that kernel density methods can be represented in terms of similarity-weighted frequencies can be found in GILBOA (2009).

and normalizing, one obtains the probability that patient  $x_0$  will show the reaction  $r$ ,

$$\Pr \{r \mid x_0\} = \frac{\sum_{x \in X} s(x_0; x) f_D(x; r)}{\sum_{r' \in R} \sum_{x \in X} s(x_0; x) f_D(x; r')} = \frac{\sum_{(x; r') \in X \times R} s(x_0; x) f_D(x; r') \hat{p}^{(x; r')}(r)}{\sum_{(x; r') \in X \times R} s(x_0; x) f_D(x; r')}, \quad (2)$$

with  $\hat{p}^{(x; r)} = \delta_r$  (the Dirichlet distributions concentrated on  $r$ ). Note that  $\hat{p}^{(x; r)}$  does not depend on the length of the database  $T$ .

Expression (2) implicitly assumes that, when collecting the data, no classification errors have occurred. Under this assumption, it is sensible to exclude all cases in which a patient has been assigned to a class different from  $r$ . Moreover, once a patient has been classified with a reaction  $r$ , the decision maker trusts that this classification is correct and assigns probability of 1 to patients of this type belonging to class  $r$ .

In practice such assumptions are hard to justify, since classifications may be biased. Suppose for example, that the data come from classifications made by different experts. Each expert observes only patients of a certain type and has the task to record their reaction. Assume that the expert dealing with class  $x$  classifies the patients correctly with probability  $1 - \epsilon_x$  and assigns them mistakenly with probability  $\frac{\epsilon_x}{|R|-1}$  to any of the remaining classes. If mistakes across experts are independent and if, in absence of better evidence, one assumes a uniform prior over the classes, then the physician will derive the following modified probability distribution,

$$\Pr \{r \mid x_0\} = \frac{\sum_{(x; r') \in X \times R} s(x_0; x) f_D(x; r') \hat{p}_T^{(x; r')}(r)}{\sum_{(x; r') \in X \times R} s(x_0; x) f_D(x; r')},$$

where the probability distributions  $\hat{p}_T^{(x; r')}$  are now posterior distributions based on  $T$  observations,

$$\begin{aligned} \hat{p}_T^{(x; r)}(r) &= \Pr \left\{ r \mid x_0; (x; r)^T \right\} = (1 - \epsilon_x)^T, \\ \hat{p}_T^{(x; r)}(r') &= \Pr \left\{ r' \mid x_0; (x; r)^T \right\} = \frac{1 - (1 - \epsilon_x)^T}{|R| - 1} \end{aligned}$$

for all  $r' \neq r$ . In this case, Axiom (A1) trivially holds. We have already shown that Bayesian updating implies Axiom (A2). Since the similarity function is independent of  $T$ , axiom (A3) applies as well. It is straightforward to check that Axiom (A4) holds, i.e.,  $\lim_{T \rightarrow \infty} \hat{p}_T^{(x; r)} = \delta_r$ . Ambiguity may become relevant if the physician is uncertain about the magnitude of the mistake

of the experts. For example, if he believes that the probability of misclassifying a type  $x$  lies in the interval  $[\underline{\epsilon}_x; \bar{\epsilon}_x] \subseteq (0; 1)$ , he immediately has to deal with a set of probability distributions

$$\hat{P}_T^{(x;r)},$$

$$\hat{P}_T^{(x;r)} = \left\{ \left( \frac{1 - (1 - \epsilon_x)^T}{|R| - 1}, \dots, \frac{1 - (1 - \epsilon_x)^T}{|R| - 1}, \underbrace{(1 - \epsilon_x)^T}_{r^{th}\text{-position}}, \frac{1 - (1 - \epsilon_x)^T}{|R| - 1}, \dots, \frac{1 - (1 - \epsilon_x)^T}{|R| - 1} \right) \mid \epsilon_x \in [\underline{\epsilon}_x; \bar{\epsilon}_x] \right\},$$

obtained by full Bayesian updating with respect to the probabilities in the set  $[\underline{\epsilon}_x; \bar{\epsilon}_x]$ . Full Bayesian updating is consistent with Axioms (A2) and (A3). Since  $\lim_{T \rightarrow \infty} \hat{P}_T^{(x;r)} = \{\delta_r\}$ , Axiom (A4) holds.

Hence, for an arbitrary database  $D$  of length  $T$ , the set of probability distributions

$$H_T(D \mid x_0) = \left\{ \frac{\sum_{(x;r') \in X \times R} s(x_0; x) f_D(x; r') \hat{P}_T^{(x;r')}}{\sum_{(x;r') \in X \times R} s(x_0; x) f_D(x; r')} \mid \hat{P}_T^{(x;r')} \in \hat{P}_T^{(x;r')} \right\} \quad (3)$$

describes the (ambiguous) beliefs of the physician about the classification of the patient of type  $x_0$  given the observations in  $D$ . Note that ambiguity vanishes, as more data become available and the physician learns the true classification scheme.

## 7.4 Constant Similarity

So far, we implicitly assumed that the similarity function is independent of the length and content of the observed data. In statistical analysis, however, the kernel width is usually chosen depending on the size of the database. The set of cases considered to be relevant for the classification of a case  $x_0$  shrinks as the number of cases increases. In our model, this corresponds to a similarity function  $s_T(x_0; x)$  which depends on  $T$ . With this modification, the representation in (3) would satisfy Axioms (A1), (A2) and (A4), but would violate Axiom (A3). Our Axioms would still allow us to determine the sets of probability distributions  $\hat{P}_T^{(x;r)}$  uniquely and to derive similarity values  $s_T(x_0; x)$  for each  $x$  and  $T$ . For small values of  $T$ , however, these similarity values would no longer be uniquely identified

Reducing the kernel width as data accumulates reflects the assumption that large databases are more representative of the distribution of  $x$ , and, therefore, contain a larger fraction of highly relevant (or even identical) cases. However, if the number of observations increases without affecting the relative frequencies of cases, then there is no reason to adjust the similarity relation

between cases. E.g., a physician with a long practice may encounter symptoms which he has never observed before. Consequently, he may find it hard to associate these new cases with those in his (long) memory. In such a situation, it does not appear reasonable to require the similarity function to converge to the identity, even for long databases. Hence, if unexpected cases are likely to occur, Axiom (A3) may be viewed as a sensible first approximation.

A non-constant similarity function may also reflect a decision maker's effort to learn about correlation between outcomes conditional on the characteristics. Consider, once again the classification problem discussed before and assume that two characteristics  $x$  and  $x'$  are very similar. If the physician has a database in which patients of types  $x$  and  $x'$  are associated with the same outcome, then this database will confirm the initial similarity perception. In contrast, if he observes that most patients of type  $x$  are classified as  $r$ , whereas most patients of type  $x'$  are classified as  $r'$ , it would be sensible to revise the similarity function. In such a case, the similarity function would depend not only on the length of the database, but also on the observed frequency of cases. Hence, both Axioms (A2) and (A3) would fail. Modelling an adjustment process of the similarity function according to the type and quantity of data is complicated by the fact that, to our knowledge, there is little systematic information in the literature about how people assess the relevance of observations.

## 7.5 Missing Data and Persistent Ambiguity

In contrast to BGSS (2004), we assume that decision makers experience ambiguity about their forecasts. GONZALES AND JAFFRAY (1998) attribute such ambiguity to missing data. In order to illustrate this possibility and how it can be incorporated in this approach, consider a physician who has to decide whether to prescribe a specific treatment  $a$  to a patient with characteristics  $x$ . He has a data-base in which the outcomes success,  $r = 1$ , or failure,  $r = 0$ , of treatment  $a$ , have been recorded for two types of patients with characteristics  $x$  and  $x'$ . Suppose that the physician has reasons to believe that the outcomes of treatment  $a$  are (perfectly) negatively correlated for patients with the two characteristics  $x$  and  $x'$  and, for simplicity, that all cases are equally relevant.

For long databases in which only outcome  $r = 1$  has been recorded for type  $x$ ,  $D = (x; a; r = 1)^T$ , it is reasonable to expect that the treatment  $a$  will lead to success,  $\hat{P}_T^{(x;a;1)} = \{\delta_1\}$ . The same prediction,  $\hat{P}_T^{(x';a;0)} = \{\delta_1\}$ , would obtain if the observed database were given by  $D' = (x'; a; r = 0)^T$ . Hence, for large databases, the probability of success can be assessed as:

$$\Pr \{r = 1 \mid D\} = f_D(x; a; 1) + f_D(x'; a; 0).$$

Suppose that for some of the observations in the database the value of the characteristic,  $x$  or  $x'$ , has not been recorded.  $\bar{x}$  is used to denote the fact that information about the characteristic is missing. In the extreme case of a database  $\bar{D} = (\bar{x}; a; r = 1)^T$  the physician cannot unambiguously determine the probability of success for the patient to be treated. Since the characteristic  $x$  or  $x'$  has not been recorded, one cannot rule out that all patients in the database  $\bar{D}$  are of type  $x$ , nor that all patients are of type  $x'$ . In the first case, he would assign a probability of 1 to a success, in the second, a probability of 0. The ignorance about the distribution because of the imprecise data can be modelled by sets of priors  $\hat{P}_T^{(\bar{x};a;1)} = \Delta^1$  and  $\hat{P}_T^{(\bar{x};a;0)} = \Delta^1$ . Combining databases in which the characteristic has been recorded with such in which data are missing gives rise to multiple priors

$$H_T(D) = \sum_{\substack{\tilde{x} \in \{x; x'; \bar{x}\} \\ r \in \{0; 1\}}} f_D(\tilde{x}; a; r) \hat{P}_T^{(\tilde{x}; a; r)}.$$

Such indeterminacy does not depend on the length of the database and will not disappear as long as there are imprecise data.

## 7.6 Selection of a Prior by the Max-Min Rule

Most of the literature on ambiguity deals with decision rules based on preferences over acts. Following ELLSBERG (1961) ambiguity aversion has become the dominant assumption about behavior under ambiguity. GILBOA AND SCHMEIDLER (1989) axiomatize a preference representation where the decision maker evaluates acts by the lowest expected utility of the act over all probabilities in a set of priors. Such conservative behavior is also recommended by the precautionary principle in environmental economics. Whatever the justification for the minimum rule, it selects a particular probability distribution from the set of priors. We will show by ex-

ample that, whenever the set of priors of a the decision maker satisfies the axioms proposed in this paper, then the probability distributions selected by the minimum principle will also obey these axioms as well<sup>12</sup>.

Reconsider the case of Example 4.2, in which the decision maker observes only draws from Urn 2. He wants to predict the outcome of a bet on a black ball drawn from this urn, action  $a_0$ . Suppose that the decision maker's beliefs are represented by a set of probability distributions as given by a correspondence  $H$  satisfying Axioms (A1) — (A5), e.g., the correspondence  $H$  in Equation (1).

Let the utility of the decision maker from a black ball been drawn from Urn 2 be 1, and 0 otherwise. Hence,  $u = (0; 1)$  is the utility vector associated with action  $a_0$ . Consider the max-min rule, which selects the single probability distribution

$$h^{\min}(D \mid x_0; a_0) = \arg \min \{h \cdot u \mid h \in H(2; a_0 \mid D)\}$$

from the set  $H(D \mid x_0; a_0)$ . If a decision maker uses this probability distribution  $h^{\min}$  to form the expected utility of  $a_0$ , then his behavior will be governed by the max-min rule, as described in GILBOA AND SCHMEIDLER (1989) and MANSKI (2009). We now demonstrate that the selection  $h^{\min}$  also satisfies our Axioms.

Since  $H$  satisfies Axiom (A1), Invariance, so does  $h^{\min}$ . Consider databases  $F, D_1 \dots D_n$  of length  $T$  such that  $D_1 \circ \dots \circ D_n = F^n$ . Then there exist positive coefficients  $\gamma_j$  with  $\sum_{j=1}^n \gamma_j = 1$  for which  $\sum_{j=1}^n \gamma_j f_{D_j} = f_F$  holds. By Axiom, (A2), we have

$$H(2; a_0 \mid F) = \sum_{j=1}^n \lambda_j H(2; a_0 \mid D_j)$$

for positive coefficients<sup>13</sup>  $\lambda_j$  such that  $\sum_{j=1}^n \lambda_j = 1$ . Note that

$$\sum_{j=1}^n \lambda_j \min \{h \cdot u \mid h \in H(2; a_0 \mid D_j)\} = \min \{h \cdot u \mid h \in H(2; a_0 \mid F)\}$$

<sup>12</sup> Similar results can be established for the rule which selects the Steiner point of each set, as well as for the more general  $\alpha$ -max-min rule, see GHIRARDATO, MACCHERONI, AND MARINACCI (2004) and CHATEAUNEUF, EICHBERGER, AND GRANT (2007).

<sup>13</sup> It appears reasonable to assume that the decision maker in this example perceives all cases to be equally relevant for his evaluation, as, e.g., in the case of Equation (1). Hence,  $\lambda_j = \gamma_j$  can be assumed for all  $j \in \{1 \dots n\}$ .

Furthermore,

$$\begin{aligned}
\sum_{j=1}^n \lambda_j \min \{h \cdot u \mid h \in H(2; a_0 \mid D_j)\} &= \sum_{j=1}^n \lambda_j (h_j^{\min} \cdot u) = \\
&= \min \{h \cdot u \mid h \in H(2; a_0 \mid F)\} \\
&= h^{\min} \cdot u,
\end{aligned}$$

where  $h_j^{\min} =: h^{\min}(2; a_0 \mid D_j)$  (note that for  $u = (0; 1)$ , these values are unique). Hence,

$$\sum_{j=1}^n \lambda_j h_j^{\min} = h^{\min}(2; a_0 \mid F),$$

implying that  $h^{\min}$  satisfies the Axiom (A2).

Since the weights  $\lambda_j$  are independent of  $T$ , Axiom (A3) is satisfied as well. For  $T \rightarrow \infty$ ,  $H_T(2; a_0 \mid (2; B)^T) \rightarrow (0; 1)$  and  $H_T(2; a_0 \mid (2; W)^T) \rightarrow (1; 0)$ . It follows that the associated probability distributions in  $h^{\min}$  will also converge to these values, implying that Learning, Axiom (A4) is satisfied. Since we have assumed only two outcomes, 0 and 1, Axiom (A5), Non-collinearity, is trivially satisfied.

## 8 Concluding Remarks

Most of the literature on ambiguity takes the degree of ambiguity as a personal subjective characteristic. In particular, there is no formal reference to the information available to the decision maker. The amount of data is, however, likely to influence both the forecast made by the decision maker and his confidence in this forecast. In this paper, we provided an approach which combines this intuition with the similarity-weighted frequency approach of BGSS (2005). We relax the Concatenation Axiom of BGSS (2005) by restricting it to databases of equal length. We show that the main result of BGSS (2005), namely that the similarity function is unique, remains true if one imposes consistency on the weights across databases of different size. This consistency is essential for the uniqueness of the similarity weights.

If one views the perception of similarity 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 database may provide not only information about the distribution

of payoffs, but also about the similarity of alternatives. One may conjecture that the more observations a database contains, the more precise the perception of similarity may become. 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. One may interpret this approach as learning similarity where similarity values are restricted to zero or one. A more general model would consider a continuum of similarity values.

A further important research question concerns the derivation from preferences of a case-based multiple-prior representation of beliefs jointly with a decision rule. 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 conditional on a database. We pursue this issue in EICHBERGER AND GUERDJIKOVA (2008).

## Appendix A. Proofs

To prove Theorem 6.1, we proceed as follows. In a first step, Lemma A.1 establishes the necessity of Axioms (A1), (A2) and (A3) for the representation.

The second step of the proof, Proposition A.2 consists in showing the result of Theorem 6.1 for the special case where predictions are single-valued, i.e. where beliefs are described by a function  $h : \mathbb{D} \rightarrow \Delta^{|R|-1}$  satisfying Axioms (A1) — (A5). Using Axioms (A4) and (A5), Lemma A.3 shows that for sufficiently large values of  $T$ , no three of the vectors  $h_T(c^T)$  are collinear. Lemma A.4 uses Axiom (A3) to demonstrate that similarity weights are independent of the length of the database. In particular, if the representation holds for a given  $D \in \mathbb{D}$ , then it holds for all databases with the same frequency  $f_D$ , regardless of their length. In Lemmas A.5 — A.7, we apply the construction of BGSS (2005) to determine the similarity function and show that the representation holds for large values of  $T$ . Lemma A.4 implies that the representation holds for all values of  $T$  if the sets  $H_T(c^T)$  are singletons  $H_T(c^T) = \{h_T(c^T)\}$ .

In the third step of the proof, we show that a correspondence  $H$  satisfying the Axioms (A1) —

(A5) can be represented as a union of functions  $h \in \mathcal{H}$  with the following properties: (i) all of these functions satisfy Axioms (A1) — (A5); (ii) for all of these functions the coefficients  $\lambda$  specified in Proposition A.2 are the same and coincide with the coefficients  $\lambda$  for the correspondence  $H$  specified in Axiom (A2); (iii) the sets  $H_T(c^T)$  are given by  $\{h_T(c^T)\}_{h \in \mathcal{H}}$ . By property (i), a representation exists for each of the functions  $h$ . Property (ii) implies that all of these representations feature the same similarity function  $s$ . Property (iii) means that we can identify  $\hat{P}_T^c$  with  $H_T(c^T)$  and use the similarity function  $s$  to obtain a representation for  $H$ .

**Lemma A.1** *The Axioms (A1), Invariance, (A2), Concatenation for databases of equal length and (A3), Constant similarity are necessary for  $H$  to have a representation of the type*

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

**Proof of Lemma A.1:**

It is obvious that for a given  $D \in \mathbb{D}^T$ ,

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

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

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

Let  $F^n = D_1 \circ \dots \circ D_n$  for some  $n \in \mathbb{N}$  and some sets  $D_1 \dots D_n \in \mathbb{D}^T$ , then

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

Hence,

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

with

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

Since  $s(c) > 0$  for all  $c \in C$ ,  $(\lambda^i)_{i=1}^n \in \text{int}(\Delta^{n-1})$  and, therefore, Axiom (A2) is satisfied.

Note further that  $\lambda$  does not depend on the length  $T$  of the databases  $D_1 \dots D_n$  and  $F$ , but only on their frequencies. Hence, if  $H_T(F^k) = \sum_{i=1}^n \lambda^i H_T(D_i^k)$  for some  $k \in \mathbb{N}$ , it holds for any  $k \in \mathbb{N}$ , implying that Axiom (A3) holds. ■

Denote by  $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$  the set of rational probability vectors of dimension  $|C|$ . The possible frequency vectors which can be generated by a database of length  $T$  are given by the set:

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

Obviously, for each  $T \in \{2, 3, \dots\}$ ,  $Q_T^{|C|} \subseteq \mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$ . Let  $\hat{f}^j$  be the  $j$ -th unit vector in  $\Delta^{|C|-1}$ .  $f(j)$  denotes the  $j$ -th component of the frequency vector  $f$ . The following proposition establishes the result of the Theorem for the special case, in which  $H$  is a function:

**Proposition A.2** *Assume that a class of functions  $h_T : Q_T^{|C|} \rightarrow \Delta^{|R|-1}$  for  $T \geq 2$  satisfies:*

(i) *for all distinct  $f, f'$  and  $f''$  such that  $\gamma f + (1 - \gamma) f' = f''$  for some  $\gamma \in (0; 1)$ , there exists a  $\lambda \in (0; 1)$  such that:*

$$\lambda h_T(f) + (1 - \lambda) h_T(f') = h_T(f'')$$

*holds for all  $T$  such that  $f, f'$  and  $f'' \in Q_T^{|C|}$ ;*

(ii) *the sequences  $\left( h_T(\hat{f}^j) \right)_{T \in \{2, 3, \dots\}}$  converge for all  $j \in \{1 \dots |C|\}$ ;*

(iii) *the set  $\left( h_\infty(\hat{f}^j) \right)_{j \in \{1 \dots |C|\}}$  contains no three collinear vectors.*

*Then, there are positive numbers  $\{s_j\}_{j=1}^{|C|}$ , which are unique up to a multiplication by a positive number, and, for each  $T \geq 2$ , there are unique probability vectors  $\{\hat{p}_T^j\}_{j=1}^{|C|}$ , such that for each  $T \geq 2$  and each  $f \in Q_T^{|C|}$ ,*

$$h_T(f) = \frac{\sum_{j=1}^{|C|} s_j f(j) \hat{p}_T^j}{\sum_{j=1}^{|C|} s_j f(j)}. \quad (\text{A-1})$$

### **Proof of Proposition A.2:**

We start with the following Lemma:

**Lemma A.3** Assume that conditions (ii) and (iii) of Proposition A.2 hold. For any three distinct  $i, j$  and  $k \in \{1 \dots |C|\}$  there exists a finite  $\bar{T}^{\{i,j,k\}}$  such that the vectors  $\left(h_T(\hat{f}^l)\right)_{l \in \{i,j,k\}}$  are non-collinear for all  $T \geq \bar{T}^{\{i,j,k\}}$ .

**Proof of Lemma A.3:**

Let  $d$  denote the distance between the point  $h_\infty(\hat{f}^i)$  and the line through the two points  $h_\infty(\hat{f}^j)$  and  $h_\infty(\hat{f}^k)$ . Since the three points are non-collinear,  $d > 0$ . Since the sequences  $h_T(\hat{f}^l)$  are converging for  $l \in \{i, j, k\}$ , we know that for each  $l$ , there exists a  $\bar{T}_l$  such that for all  $T \geq \bar{T}_l$ ,  $h_T(\hat{f}^l)$  is contained in a ball with a center in  $h_\infty(\hat{f}^l)$  and with a radius  $\frac{d}{3}$ , denoted by  $h_T(\hat{f}^l) \in B_{h_\infty(\hat{f}^l)}(d/3)$ . Let  $\bar{T}^{\{i,j,k\}} =: \max_{l \in \{i,j,k\}} \{\bar{T}_l\}$ . Take any two points  $x^j \in B_{h_\infty(\hat{f}^j)}(d/3)$  and  $x^k \in B_{h_\infty(\hat{f}^k)}(d/3)$  and note that the line which connects these two points must be at a distance at least  $\frac{d}{3}$  from any point  $x^i \in B_{h_\infty(\hat{f}^i)}(d/3)$ . Hence,  $x^i, x^j$  and  $x^k$  cannot be collinear. Since for every  $T \geq \bar{T}^{\{i,j,k\}}$ , and every  $l \in \{i, j, k\}$ ,  $h_T(\hat{f}^l) \in B_{h_\infty(\hat{f}^l)}(d/3)$ , the three vectors cannot be collinear. ■

Let

$$\bar{T} =: \max_{\{\{i,k,l\} \subseteq C\}} \{\bar{T}^{\{i,k,l\}}\}$$

For each  $T$ , define  $\hat{p}_T^j =: h_T(\hat{f}^j)$ . Since each of the sequences  $\left(h_T(\hat{f}^j)\right)_{T \in \{2,3,\dots\}}$  converges and no three limit vectors are collinear, the sequences  $(\hat{p}_T^j)_{T \in \{2,3,\dots\}}$  inherit these properties.

We have to show that there are positive numbers  $\{s_j\}_{j=1}^{|C|}$  such that (A-1) holds. The next Lemma demonstrates that if such weights can be used to represent  $h_T(f)$  for some  $f \in Q_T^{|C|}$ , then the same weights can be used to represent  $h_{T'}(f)$  for any  $T'$  such that  $f \in Q_{T'}^{|C|}$ .

**Lemma A.4** For  $|C| \geq 3$ , let  $\{s_j\}_{j=1}^{|C|}$  be a collection of similarity weights. For any  $T \geq 2$  and any  $f \in Q_T^{|C|}$ , define the function  $g_T(f)$  by

$$g_T(f) = \frac{\sum_{j=1}^{|C|} s_j f(j) \hat{p}_T^j}{\sum_{j=1}^{|C|} s_j f(j)}$$

Suppose that for some  $f \in Q_T^{|C|}$ , we can show that  $h_T(f) = g_T(f) = \frac{\sum_{j=1}^{|C|} s_j f(j) \hat{p}_T^j}{\sum_{j=1}^{|C|} s_j f(j)}$ . Then,  $h_{T'}(f) = g_{T'}(f)$  for all  $T'$  such that  $f \in Q_{T'}^{|C|}$ .

**Proof of Lemma A.4:**

Let  $T_f$  be the smallest  $T$  such that  $f \in Q_T^{|C|}$ . Then  $f \in Q_T^{|C|}$  iff  $T = kT_f$  for some  $k \in \mathbb{N}$ . If

$\lambda \in \Delta^{|C|-1}$  with  $\lambda_j = 0$  iff  $f(j) = 0$  satisfies

$$h_{T_f}(f) = \sum_{j=1}^{|C|} \lambda_j h_{T_f}(\hat{f}^j) = \sum_{j=1}^{|C|} \lambda_j \hat{p}_{T_f}^j,$$

then, by property (i) in Proposition A.2,

$$h_{kT_f}(f) = \sum_{j=1}^{|C|} \lambda_j h_{kT_f}(\hat{f}^j) = \sum_{j=1}^{|C|} \lambda_j \hat{p}_{kT_f}^j.$$

In particular,

$$h_T(f) = \sum_{j=1}^{|C|} \lambda_j h_T(\hat{f}^j) = \sum_{j=1}^{|C|} \lambda_j \hat{p}_T^j = \frac{\sum_{j=1}^{|C|} s_j f(j) \hat{p}_T^j}{\sum_{j=1}^{|C|} s_j f(j)} = g_T(f),$$

implying  $\lambda_i = \frac{s_i f(i)}{\sum_{j=1}^{|C|} s_j f(j)}$  for  $i \in \{1, 2, \dots, |C|\}$ . Since  $g_{kT_f}(f) = \frac{\sum_{j=1}^{|C|} s_j f(j) \hat{p}_{kT_f}^j}{\sum_{j=1}^{|C|} s_j f(j)} = \sum_{j=1}^{|C|} \lambda_j \hat{p}_{kT_f}^j = h_{kT_f}(f)$ , we have the desired result. ■

We now prove the result of Proposition A.2 for the case  $|C| = 3$ . For this case, define  $f^* =: \sum_{j=1}^3 \frac{1}{3} \hat{f}^j$  and consider  $T = 3\bar{T}$ . Obviously,  $f^* \in Q_{3\bar{T}}^3$ . Let  $s_1, s_2$  and  $s_3$  be the unique up to a multiplication by a positive number solution of the equation:

$$h_{3\bar{T}}(f^*) = \frac{\sum_{j=1}^3 s_j \hat{p}_{3\bar{T}}^j}{\sum_{j=1}^3 s_j}.$$

For any  $T \geq 2$  and any  $f \in Q_T^3$ , define  $g_T(f) =: \frac{\sum_{j=1}^3 s_j f(j) \hat{p}_T^j}{\sum_{j=1}^3 s_j f(j)}$ . Obviously,  $g_{3\bar{T}}(f^*) = h_{3\bar{T}}(f^*)$ .

Lemma A.4 then implies

$$g_{3k}(f^*) = h_{3k}(f^*)$$

for all  $k \in \mathbb{N}$ .

In order to state our next Lemma, we define the first simplicial partition of  $\mathbb{Q}_+^{|C|} \cap \Delta^{|C|-1}$  by the four triangles separated by the segments connecting the three points  $\left(\frac{1}{2}\hat{f}^i + \frac{1}{2}\hat{f}^j\right)$  for  $i \neq j$ . The second simplicial partition is obtained by partitioning each of the simplicial triangles into simplicial triangles, the  $k$ -th simplicial partition is defined recursively. The simplicial points of the  $k$ -th simplicial partition are all the vertices of the triangles in this partition. Note that all elements of the first simplicial partition are in  $Q_2^3$ . The centers of gravity of these triangles are in  $Q_6^3$ . All elements of the second simplicial partition are in  $Q_4^3$ , while their centers of gravity

are in  $Q_{12}^3$ , etc. Our next result shows that, for  $|C| = 3$ , the functions  $h$  and  $g$  coincide on the set of simplicial points.

**Lemma A.5** *Let  $\text{conv}(\hat{f}_k^1, \hat{f}_k^2, \hat{f}_k^3)$  be a simplicial triangle from the  $k$ -th simplicial partition and let  $f_k^* = \sum_{i=1}^3 \frac{1}{3} \hat{f}_k^i$  be its center of gravity. Let  $T_0 \geq \bar{T}$  be such that  $\{\hat{f}_k^i\}_{i=1}^3$  and  $f_k^* \in Q_{T_0}^3$ . If  $h_{T_0}(\hat{f}_k^i) = g_{T_0}(\hat{f}_k^i)$  for all  $i = 1 \dots 3$  and  $h_{T_0}(f_k^*) = g_{T_0}(f_k^*)$ , then, for every simplicial triangle  $\text{conv}(\hat{f}_l^1, \hat{f}_l^2, \hat{f}_l^3)$  of  $\text{conv}(\hat{f}_k^1, \hat{f}_k^2, \hat{f}_k^3)$ , with a center of gravity  $f_l^* = \sum_{i=1}^3 \frac{1}{3} \hat{f}_l^i$ ,  $h_{12T_0}(\hat{f}_l^i) = g_{12T_0}(\hat{f}_l^i)$  for all  $i = 1 \dots 3$  and  $h_{12T_0}(f_l^*) = g_{12T_0}(f_l^*)$  holds.*

**Proof of Lemma A.5:**

Observe that if  $(f_1, f_2)$  and  $(f_3, f_4)$  are non-collinear segments in  $Q_{T_0}^3$  with the property that  $h_{T_0}(f_i) = g_{T_0}(f_i)$  for all  $i \in \{1 \dots 4\}$ , then for  $f = (f_1; f_2) \cap (f_3; f_4)$ ,  $h_{T_{0f}}(f) = g_{T_{0f}}(f)$ , where  $T_{0f} = \min\{T \geq T_0 \mid f \in Q_T^3\}$ .

In particular, let  $\text{conv}(\hat{f}_k^1, \hat{f}_k^2, \hat{f}_k^3)$  be a simplicial triangle from the  $k$ -th simplicial partition and let  $h_{T_0}(\hat{f}_k^i) = g_{T_0}(\hat{f}_k^i)$  for all  $i = 1 \dots 3$  and  $h_{T_0}(f_k^*) = g_{T_0}(f_k^*)$ . By Lemma A.4,  $h_{12T_0}(\hat{f}_k^i) = g_{12T_0}(\hat{f}_k^i)$  for all  $i = 1 \dots 3$  and  $h_{12T_0}(f_k^*) = g_{12T_0}(f_k^*)$ . Note that for any two  $i, j \in \{1, 2, 3\}$ ,  $i \neq j$  and  $n = \{1; 2; 3\} \setminus \{i; j\}$ ,

$$\frac{1}{2}\hat{f}_k^i + \frac{1}{2}\hat{f}_k^j = (\hat{f}_k^i; \hat{f}_k^j) \cap (\hat{f}_k^n; f_k^*).$$

Since  $h_{12T_0}(\hat{f}_k^1)$ ,  $h_{12T_0}(\hat{f}_k^2)$  and  $h_{12T_0}(\hat{f}_k^3)$  are non-collinear, and, hence,  $(h_{12T_0}(\hat{f}_k^i); h_{12T_0}(\hat{f}_k^j))$  and  $(h_{12T_0}(\hat{f}_k^n); h_{12T_0}(f_k^*))$  are non-collinear as well, they have a unique intersection point.

Since both  $g_{12T_0}(\frac{1}{2}\hat{f}_k^i + \frac{1}{2}\hat{f}_k^j)$  and  $h_{12T_0}(\frac{1}{2}\hat{f}_k^i + \frac{1}{2}\hat{f}_k^j)$  must coincide with this intersection point, it follows that  $g_{12T_0}(\frac{1}{2}\hat{f}_k^i + \frac{1}{2}\hat{f}_k^j) = h_{12T_0}(\frac{1}{2}\hat{f}_k^i + \frac{1}{2}\hat{f}_k^j)$ .

Now consider the centers of gravity of the four subtriangles. For the triangle

$$\text{conv}\left(\frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^2, \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3, \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3\right)$$

the center of gravity is  $f_k^*$  and, hence, satisfies the condition. Consider, therefore, w.l.o.g., the triangle  $\text{conv}(\hat{f}_l^1, \hat{f}_l^2, \hat{f}_l^3)$  with  $\hat{f}_l^1 = \hat{f}_k^3$ ,  $\hat{f}_l^2 = \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3$  and  $\hat{f}_l^3 = \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3$ . First note that since  $(\frac{1}{2}(\frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3) + \frac{1}{2}(\frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3))$  is the intersection of  $(\hat{f}_k^3; \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^2)$  and

$\left(\frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3, \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3\right)$ , we have:

$$h_{12T_0} \left( \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3 \right) + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3 \right) \right) = g_{12T_0} \left( \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3 \right) + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3 \right) \right).$$

Similarly,

$$h_{12T_0} \left( \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3 \right) + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^2 \right) \right) = g_{12T_0} \left( \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3 \right) + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^2 \right) \right)$$

The point  $\frac{1}{2}\hat{f}_k^3 + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3 \right) = \frac{3}{4}\hat{f}_k^3 + \frac{1}{4}\hat{f}_k^2$  is on the intersection of

$$\left( \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3 \right) + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3 \right) \right); \left( \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3 \right) + \frac{1}{2} \left( \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^2 \right) \right)$$

and  $\left(\hat{f}_k^2, \hat{f}_k^3\right)$ . Hence,  $h_{12T_0} \left( \frac{3}{4}\hat{f}_k^3 + \frac{1}{4}\hat{f}_k^2 \right) = g_{12T_0} \left( \frac{3}{4}\hat{f}_k^3 + \frac{1}{4}\hat{f}_k^2 \right)$ . The center of gravity  $f_l^*$  of

$$\text{conv} \left( \hat{f}_k^3, \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3, \frac{1}{2}\hat{f}_k^2 + \frac{1}{2}\hat{f}_k^3 \right)$$

is the intersection of  $\left(\hat{f}_k^3, \frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^2\right)$  and  $\left(\left(\frac{1}{2}\hat{f}_k^1 + \frac{1}{2}\hat{f}_k^3\right); \left(\frac{3}{4}\hat{f}_k^3 + \frac{1}{4}\hat{f}_k^2\right)\right)$ , and, hence,  $h_{12T_0}(f_l^*) = g_{12T_0}(f_l^*)$ . ■

Applying the claim inductively and using the result of Lemma A.4, we conclude that the functions  $h$  and  $g$  coincide on the set of all simplicial points.

To complete the proof of Proposition A.2 for the case of  $|C| = 3$ , it remains to show that the functions  $h$  and  $g$  coincide on the set of all rational points.

**Lemma A.6**  $h_T(f) = g_T(f)$  for all  $f \in Q_T^3$  and for all  $T \geq 2$ .

### Proof of Lemma A.6:

Take an arbitrary  $f \in Q_T^3$ . For some  $\hat{T} \geq \bar{T}$ , let  $f \in Q_{\hat{T}}^3$ . Form a sequence  $T^1 \dots T^k \dots$  with  $T^1 = 6\hat{T}$  and  $T^k = 6T^{k-1}$  and take a sequence of simplicial triangles  $\left(\hat{f}_1^1; \hat{f}_1^2; \hat{f}_1^3\right) \dots \left(\hat{f}_k^1; \hat{f}_k^2; \hat{f}_k^3\right) \dots$  such that for each  $k$ ,  $\hat{f}_k^1$ ,  $\hat{f}_k^2$  and  $\hat{f}_k^3$  are in  $Q_{T^k}^3$ ,  $f \in \text{conv} \left( \hat{f}_k^1, \hat{f}_k^2, \hat{f}_k^3 \right)$  and  $\lim_{k \rightarrow \infty} \hat{f}_k^i = f$  for all  $i \in \{1, 2, 3\}$ . It is obvious that this construction is possible for every  $f$ . We want to show that  $h_{T^k}(f) = g_{T^k}(f)$  for all  $k$ .

First note that if  $f \in \text{conv} \left( \hat{f}_k^1; \hat{f}_k^2; \hat{f}_k^3 \right)$ , then by the definition of  $g$ ,

$$g_{T^k}(f) \in \text{conv} \left( g_{T^k} \left( \hat{f}_k^1 \right); g_{T^k} \left( \hat{f}_k^2 \right); g_{T^k} \left( \hat{f}_k^3 \right) \right).$$

Since for all  $n \in \{1, 2, 3\}$ ,  $\lim_{k \rightarrow \infty} \hat{p}_{T^k}^n = h_\infty(\hat{f}^n) \in \Delta^2$ , we have that for all  $r \in R$  and all

$j \in \{1, 2, 3\}$

$$\begin{aligned}
& \lim_{k \rightarrow \infty} \left( \frac{\sum_{n=1}^3 s_n f(n) \hat{p}_{T^k}^n(r)}{\sum_{n=1}^3 s_n f(n)} - \frac{\sum_{n=1}^3 s_n \hat{f}_k^j(n) \hat{p}_{T^k}^n(r)}{\sum_{n=1}^3 s_n \hat{f}_k^j(n)} \right) \\
&= \lim_{k \rightarrow \infty} \sum_{n=1}^3 \hat{p}_{T^k}^n(r) \left( \frac{s_n f(n)}{\sum_{n=1}^3 s_n f(n)} - \frac{s_n \hat{f}_k^j(n)}{\sum_{n=1}^3 s_n \hat{f}_k^j(n)} \right) \\
&= \sum_{n=1}^3 h_\infty(\hat{f}^n)(r) \left( \frac{\sum_{i=1}^3 s_i f(i)}{\sum_{n=1}^3 s_n f(n)} - \frac{\sum_{i=1}^3 s_i \hat{f}^j(i)}{\sum_{n=1}^3 s_n \hat{f}^j(n)} \right) = 0,
\end{aligned}$$

where  $\hat{p}_{T^k}^n(r)$  and  $h_\infty(\hat{f}^n)(r)$  denote the  $r^{\text{th}}$ -components of the vectors  $\hat{p}_{T^k}^n$  and  $h_\infty(\hat{f}^n)$ , respectively. Hence,

$$\lim_{k \rightarrow \infty} \|g_{T^k}(f) - g_{T^k}(\hat{f}_k^j)\| = 0$$

for all  $j \in \{1, 2, 3\}$ .

Property (i) in Proposition A.2 implies that there exists a vector  $\lambda \in \Delta^2$ , independent of  $T^k$ , such that

$$h_{T^k}(f) = \sum_{n=1}^3 \lambda_n h_{T^k}(\hat{f}^n).$$

Suppose that for some  $i \in \{1, 2, 3\}$ ,  $\lambda_i \neq \frac{s_i f(i)}{\sum_{n=1}^3 s_n f(n)}$ , i.e. that there exists an  $\epsilon > 0$  such that:

$$\left\| \lambda_i - \frac{s_i f(i)}{\sum_{n=1}^3 s_n f(n)} \right\| = \epsilon > 0.$$

Since  $h_{T^k}(f) \in \text{conv}(h_{T^k}(\hat{f}^1); h_{T^k}(\hat{f}^2); h_{T^k}(\hat{f}^3))$  and

$$\text{conv}(h_{T^k}(\hat{f}^1); h_{T^k}(\hat{f}^2); h_{T^k}(\hat{f}^3)) = \text{conv}(g_{T^k}(\hat{f}^1); g_{T^k}(\hat{f}^2); g_{T^k}(\hat{f}^3))$$

we have that

$$\lim_{k \rightarrow \infty} \text{conv}(h_{T^k}(\hat{f}^1); h_{T^k}(\hat{f}^2); h_{T^k}(\hat{f}^3)) \setminus \{g_{T^k}(f)\} = \emptyset.$$

Hence, it must be that:

$$\lim_{k \rightarrow \infty} \|h_{T^k}(f) - g_{T^k}(f)\| = 0,$$

which is equivalent to:

$$\lim_{k \rightarrow \infty} \left\| \sum_{i=1}^3 \left( \lambda_i - \frac{s_i f(i)}{\sum_{n=1}^3 s_n f(n)} \right) \hat{p}_{T^k}^i \right\| = 0, \quad (\text{A-2})$$

which reduces to

$$\left\| \sum_{i=1}^3 \left( \lambda_i - \frac{s_i f(i)}{\sum_{n=1}^3 s_n f(n)} \right) h_\infty(\hat{f}^i) \right\| = 0, \quad (\text{A-3})$$

By presumption, the vectors  $\left(h_\infty \left(\hat{f}^i\right)\right)_{i \in \{1,2,3\}}$  are not collinear. Hence, (A-3) can be satisfied only if  $\lambda_i = \frac{s_i f(i)}{\sum_{n=1}^3 s_n f(n)}$  for all  $i \in \{1, 2, 3\}$ . It follows that  $h_{T^k}(f) = g_{T^k}(f)$  for all  $k$ . Lemma A.4 then implies that  $h_T(f) = g_T(f)$  for all  $T$  such that  $f \in Q_T^3$ . ■

Lemma A.6 completes the proof of Proposition A.2 for the case of  $|C| = 3$ .

Now consider the case of  $|C| > 3$ . To define the similarity function for this case, choose three distinct  $j, k, l \leq |C|$ . Define  $f^{\{j,k,l\}} =: \sum_{n \in \{j,k,l\}} \frac{1}{3} \hat{f}^n$  and let  $s^{\{j,k,l\}}$  be the unique (up to a multiplication by a positive number) solution of:

$$h_{3\bar{T}}(f^{\{j,k,l\}}) = \frac{\sum_{n \in \{j,k,l\}} s_n^{\{j,k,l\}} \hat{p}_{3\bar{T}}^n}{\sum_{n \in \{j,k,l\}} s_n^{\{j,k,l\}}}.$$

Define

$$g_T^{\{j,k,l\}}(f) =: \frac{\sum_{n \in \{j,k,l\}} s_n^{\{j,k,l\}} f(n) \hat{p}_T^n}{\sum_{n \in \{j,k,l\}} s_n^{\{j,k,l\}} f(n)}.$$

for all  $T$  and all  $f \in Q_T^3 \cap \text{conv} \left\{ \hat{f}^j, \hat{f}^k, \hat{f}^l \right\}$ . Note that  $g_T^{\{j,k,l\}} = h_T$  on this set. We first show that the similarity values  $s_n^{\{j,k,l\}}$  determined in this way do not depend on the choice of  $j, k$  and  $l$ : note that  $\frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} = \frac{s_j^{\{j,k,l'\}}}{s_k^{\{j,k,l'\}}}$ , since  $g_T^{\{j,k,l\}}(f) = g_T^{\{j,k,l'\}}(f) = h_T(f)$  for all  $f \in \text{conv} \left( \hat{f}^j, \hat{f}^k \right)$ .

Hence, define

$$\gamma_{jk} =: \frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} = \frac{s_j^{\{j,k,l'\}}}{s_k^{\{j,k,l'\}}}$$

for all  $l, l' \in \{1 \dots |C|\}$ . Note that  $\gamma_{jk}$  is well-defined for all  $j, k \leq |C|$ , since no three vectors  $\hat{p}_{3\bar{T}}^j = h_{3\bar{T}}(\hat{f}^j)$ ,  $\hat{p}_{3\bar{T}}^k = h_{3\bar{T}}(\hat{f}^k)$  and  $\hat{p}_{3\bar{T}}^l = h_{3\bar{T}}(\hat{f}^l)$  are collinear. Furthermore,

$$\gamma_{jk} \gamma_{kl} \gamma_{lj} = \frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} \frac{s_k^{\{j,k,l\}}}{s_l^{\{j,k,l\}}} \frac{s_l^{\{j,k,l\}}}{s_j^{\{j,k,l\}}} = 1.$$

Define  $s_1 =: 1$  and  $s_j =: \gamma_{j1}$  for all  $j \in \{2 \dots |C|\}$ . We wish to show that for all triples  $j, k$  and  $l$ ,  $\left\{ s_n^{\{j,k,l\}} \right\}_{n \in \{j,k,l\}}$  is proportional to  $\{s_j; s_k; s_l\}$ . This follows from  $\gamma_{1j} \gamma_{jk} \gamma_{k1} = \frac{1}{s_j} \frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} s_k = 1$ . Given  $s = (s_j)_{j=1}^{|C|}$ , define

$$g_T(f) =: \frac{\sum_{j=1}^{|C|} s_j f(j) \hat{p}_T^j}{\sum_{j=1}^{|C|} s_j f(j)}$$

for all  $T$  and all  $f \in Q_T^{|C|}$ .

We know that for  $|C| = 3$ ,  $g_T(f) = h_T(f)$ . We wish to prove that the same is true for any  $|C| \geq 3$ . We proceed by induction. Suppose that the claim is true for all  $|C| \leq N$  and take  $|C| = N + 1$ . We prove the following claim by induction:

**Lemma A.7** For every subset  $K \subseteq \{1 \dots |C|\}$ ,  $h_T(f) = g_T(f)$  holds for every  $T \geq 2$  and every  $f \in Q_T^{|K|} \cap \text{conv} \left( \left\{ \hat{f}^j \mid j \in K \right\} \right)$ .

**Proof of Lemma A.7:**

We know that the claim is true for  $|K| = 3$ , so we assume that it is true for  $|K| = N$  and prove that it will hold for  $|K| = N + 1$ .

By property (iii) of Proposition A.2, no three of the vectors  $h_\infty(\hat{f}^i)$  are collinear. By the induction argument, for every  $m \in K$ ,  $h_T(f) = g_T(f)$  holds for every  $T \geq 2$  and every  $f \in Q_T^N \cap \text{conv} \left( \left\{ \hat{f}^j \mid j \in K \setminus \{m\} \right\} \right)$ .

Let  $T \geq \bar{T}$  and consider first  $f \in \text{int} \left( Q_T^{N+1} \cap \text{conv} \left( \left\{ \hat{f}^j \mid j \in K \right\} \right) \right)$ .  $f$  can be expressed as:  $f = \sum_{l=1}^{N+1} \gamma_l \hat{f}^l$  for some rational coefficients  $\gamma_l > 0$ . For every  $m \in K$ , let  $f_m$  be the point in

$$\text{conv} \left( \left( \left\{ \hat{f}^l \mid l \in K \setminus m \right\} \right) \cap Q_{T'}^N \right)$$

that is on the line connecting  $f$  and  $\hat{f}^m$ , i.e.  $f_m = \sum_{\substack{l=1 \\ l \neq m}}^{N+1} \frac{\gamma_l}{1-\gamma_m} \hat{f}^l$ , where  $T'$  is the smallest number of observations, for which  $f$  and  $f_m \in Q_{T'}^N$  for all  $m \in K$ . Such a  $T'$  exists, since  $\gamma_l$  are rational coefficients. We have  $h_{T'}(\hat{f}^m) = g_{T'}(\hat{f}^m)$  and, by the induction argument,  $h_{T'}(f_m) = g_{T'}(f_m)$ . Property (i) of Proposition A.2 implies

$$\begin{aligned} h_{T'}(f) &\in \left( h_{T'}(\hat{f}^m); h_{T'}(f_m) \right) \\ h_{T'}(f) &\in \left( h_{T'}(\hat{f}^{m'}); h_{T'}(f_{m'}) \right) \end{aligned}$$

for all  $m$  and  $m' \in K$ .

Now we wish to show that not all of these intervals are collinear. This follows from the fact that no three of the vectors  $h_\infty(\hat{f}^m)$  are collinear, and hence, by Lemma A.3, the corresponding vectors  $h_{T'}(\hat{f}^m)$  are also non-collinear for any  $T' \geq T$ . Hence, there are two distinct intervals  $\left( h_{T'}(\hat{f}^m); h_{T'}(f_m) \right)$  and  $\left( h_{T'}(\hat{f}^{m'}); h_{T'}(f_{m'}) \right)$  which do not lie on the same line, and, by property (i) of Proposition A.2, have  $h_{T'}(f)$  as an intersection point. Since,  $g_{T'}(f)$  is by construction also an intersection point of the two intervals, it follows that  $h_{T'}(f) = g_{T'}(f)$ .

By Lemma A.4, we conclude that  $h_T(f) = g_T(f)$  holds for all  $T \geq 2$  and all

$$f \in \text{int} \left( \text{conv} \left( \left\{ \hat{f}^j \mid j \in K \right\} \right) \cap Q_T^{N+1} \right),$$

as well as for all  $f \in \text{conv} \left( \left\{ \hat{f}^j \mid j \in K \setminus m \right\} \right) \cap Q_T^N$  for  $m \in K$ , thus establishing the result. ■

Lemma A.7 completes the proof of Proposition A.2.

**Lemma A.8** *Under Axiom (A5), it is possible to select vectors  $h_\infty(c) \in H_\infty(c)$  for each  $c \in C$  such that no three vectors in the set  $(h_\infty(c))_{c \in C}$  are collinear. Furthermore, for each  $c$ , there exists a  $T_c \in \mathbb{N}$  and a sequence of vectors  $\hat{h}_T(c^T) \in H_T(c)^T$  for  $T \geq T_c$  such that*

$$\lim_{T \rightarrow \infty} \hat{h}_T(c^T) = h_\infty(c).$$

**Proof of Lemma A.8:**

Denote the set  $\tilde{C}_p$  to be the set of all cases  $c \in C$ , such that  $H_\infty(c)$  is of dimension  $p \in \{0, 1, \dots, |R| - 1\}$ . To show that a selection of vectors  $h_\infty(c)$  with the stated properties exists, first set  $(h_\infty(c))$  to be the unique elements of each of the sets  $(H_\infty(c))_{c \in \tilde{C}_0}$ . No three of these are collinear by Axiom (A5). Take a case  $\hat{c} \in \tilde{C}_1$ . For a given segment  $(e, f)$ , define  $l(e, f)$  to be the line containing the segment. Consider the set

$$\mathcal{L}_{\hat{c}} = \{l(h_\infty(c'), h_\infty(c''))\}_{c', c'' \in \tilde{C}_0} \cup (l(H_\infty(c))_{c \in \tilde{C}_1} \setminus l(H_\infty(\hat{c})))$$

This is the set of all lines connecting any two singleton sets, as well as the collection of lines defined by the segments in  $\{H_\infty(c)\}_{c \in C}$ , excluding the line containing  $H_\infty(\hat{c})$  itself. Choose a point  $h_\infty(\hat{c}) \in H_\infty(\hat{c})$  such that  $h_\infty(\hat{c}) \notin \mathcal{L}_{\hat{c}}$ . That this can be done is ensured by Axiom (A5). We now show that no three of the sets  $\{h_\infty(c)\}$  for  $c \in \tilde{C}_0 \cup \{\hat{c}\}$  and  $H_\infty(c)$  for  $c \in \tilde{C}_1$  are collinear. First consider the combination of  $h_\infty(\hat{c})$  with any two points  $h_\infty(c')$  and  $h_\infty(c'')$  with  $c', c'' \in \tilde{C}_0$ . Since  $h_\infty(\hat{c}) \notin l(h_\infty(c'), h_\infty(c''))$ , these are non-collinear.

Second, consider the combination of  $h_\infty(\hat{c})$  with a point  $h_\infty(c')$ , ( $c' \in \tilde{C}_0$ ), and a segment,  $H_\infty(c'')$ , ( $c'' \in \tilde{C}_1$ ). If  $h_\infty(c')$  and  $H_\infty(c'')$  are collinear, then  $h_\infty(\hat{c}) \notin l(H_\infty(c''))$  and, hence, the three sets are not collinear. If  $h_\infty(c')$  and  $H_\infty(c'')$  are not collinear, then neither is the triple  $h_\infty(\hat{c})$ ,  $h_\infty(c')$  and  $H_\infty(c'')$ .

Last, consider the combination of  $h_\infty(\hat{c})$  with two segments  $H_\infty(c')$  and  $H_\infty(c'')$  ( $c', c'' \in \tilde{C}_1$ ). Axiom (A5) excludes the case in which all three of the sets  $H_\infty(\hat{c})$ ,  $H_\infty(c')$  and  $H_\infty(c'')$  lie

on the same line. Hence, at least one of these two segments, say  $H_\infty(c')$  is non-collinear to  $H_\infty(\hat{c})$ . It follows that  $h_\infty(\hat{c}) \notin l(H_\infty(c'))$ , which proves that the three sets  $h_\infty(\hat{c})$ ,  $H_\infty(c')$  and  $H_\infty(c'')$  are non-collinear. We have thus shown that the new set of limit predictions defined by:

$$\begin{aligned} H_\infty^1(c) &= \{h_\infty(c)\} \text{ for all } c \in \tilde{C}_0 \cup \{\hat{c}\} \\ H_\infty^1(c) &= H_\infty(c) \text{ for all } c \notin \tilde{C}_0 \cup \{\hat{c}\} \end{aligned}$$

satisfies Axiom (A5).

Using the same argument by induction, it is possible to choose points  $h_\infty(c)$  for all  $c \in \tilde{C}_0 \cup \tilde{C}_1$  in such a way that no three of these points are collinear. This procedure generates a new set of limit predictions:

$$\begin{aligned} H_\infty^{|\tilde{C}_1|}(c) &= \{h_\infty(c)\} \text{ for all } c \in \tilde{C}_0 \cup \{\tilde{C}_1\} \\ H_\infty^{|\tilde{C}_1|}(c) &= H_\infty(c) \text{ for all } c \notin \tilde{C}_0 \cup \{\tilde{C}_1\} \end{aligned}$$

Using the singleton sets  $\left(H_\infty^{|\tilde{C}_1|}(c)\right)_{c \in \tilde{C}_0 \cup \tilde{C}_1}$  defined in this way, construct the set of all lines connecting these points:

$$\mathcal{L}_{\tilde{C}_0 \cup \tilde{C}_1} = \left\{ l\left(H_\infty^{|\tilde{C}_1|}(c'), H_\infty^{|\tilde{C}_1|}(c'')\right) \right\}_{c', c'' \in \tilde{C}_0 \cup \tilde{C}_1} = \{l(h_\infty(c'), h_\infty(c''))\}_{c', c'' \in \tilde{C}_0 \cup \tilde{C}_1}.$$

Note that the intersection of each of the remaining  $H_\infty^{|\tilde{C}_1|}(\hat{c})$  with  $\mathcal{L}_{\tilde{C}_0 \cup \tilde{C}_1}$  is a finite collection of points and segments. Fix a case  $\bar{c} \notin \tilde{C}_0 \cup \tilde{C}_1$ . Since the set  $H_\infty^{|\tilde{C}_1|}(\bar{c})$  is of dimension 2 or higher, we can find a point  $h_\infty(\bar{c}) \in H_\infty^{|\tilde{C}_1|}(\bar{c}) \setminus \mathcal{L}_{\tilde{C}_0 \cup \tilde{C}_1}$ , which is non-collinear to any pair of vectors in  $(h_\infty(c))_{c \in \tilde{C}_0 \cup \tilde{C}_1}$ . Analogously, define the set  $\mathcal{L}_{\tilde{C}_0 \cup \tilde{C}_1 \cup \{\bar{c}\}} = l(h_\infty(c'), h_\infty(c''))_{c', c'' \in \tilde{C}_0 \cup \tilde{C}_1 \cup \{\bar{c}\}}$  and proceed by induction to complete the construction.

The fact that a convergent sequence  $\left(\hat{h}_T(c^T)\right)_{T \geq T_c}$  exists follows directly from the definition of the limit of a sequence of sets, see ROCKAFELLAR AND WETS (2004, p. 109). ■

### **Proof of Theorem 6.1:**

We show that we can represent the correspondence  $H$  as a collection of functions  $\mathcal{H} =:$

$\{h : \mathbb{D} \rightarrow \Delta^{|R|-1}\}$  satisfying the following conditions:

(i) for any  $T$  and any three databases  $D, D'$  and  $D'' \in \mathbb{D}^T$  such that  $\gamma f_D + (1 - \gamma) f_{D'} = f_{D''}$

for some  $\gamma \in (0; 1)$ ,

$$\lambda H_T(D) + (1 - \lambda) H_T(D') = H_T(D'')$$

implies

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

for every  $h \in \mathcal{H}$ ;

(ii) for each  $h \in \mathcal{H}$  and  $c \in C$ , the sequence  $(h_T(c^T))_{T \geq 2}$  converges to some limit  $h_\infty(c)$ ;

(iii) for each  $h \in \mathcal{H}$ , no three of the vectors in the set  $(h_\infty(c))_{c \in C}$  are collinear;

(iv) for each  $T \geq 2$ , and any  $D \in \mathbb{D}^T$ ,  $\cup_{h \in \mathcal{H}} h_T(D) = H_T(D)$ .

To construct the set of functions  $\mathcal{H}$ , proceed as follows: fix a  $D \in \mathbb{D}$ . Let  $T$  be the length of the database  $D$ . Let  $h_T(D) \in H_T(D)$ . If  $D = (c)^T$  for some  $c \in C$ , pick an element  $h_T((c')^T) \in H_T((c')^T)$  for all  $c' \neq c$ . For any  $D' \in \mathbb{D}^T$ , let

$$h_T^D(D') = \sum_{c \in |C|} \lambda'_c h_T(c^T),$$

where  $\lambda'_c$  are such that:

$$H_T(D') = \sum_{c \in |C|} \lambda'_c H_T(c^T). \quad (\text{A-4})$$

If  $D \neq c^T$  for all  $c \in C$ , let  $(\lambda_c)_{c \in C}$  be such that:

$$H_T(D) = \sum_{c \in |C|} \lambda_c H_T(c^T).$$

It is then possible to choose vectors  $\in H_T(c^T)$  for each  $c \in C$  such that:

$$h_T(D) = \sum_{c \in |C|} \lambda_c h_T(c^T).$$

Using the so chosen vectors  $h_T(c^T)$ , construct  $h_T(D')$  as in (A-4) for all other sets  $D' \in \mathbb{D}^T$ . Let  $T_C = \max\{T_c \mid c \in C\}$ , where  $T_c$  are as defined in the statement of Lemma A.8. For  $T' < T_C$ ,  $T' \neq T$ , pick any vectors  $h_{T'}(c^{T'}) \in H_{T'}(c^{T'})$  and for any  $D' \in \mathbb{D}_{T'}$ , construct the vectors  $h_{T'}(D')$  as in (A-4).

To complete the construction for  $T' \geq T_C$ ,  $T' \neq T$ , choose vectors  $\hat{h}_\infty(c) \in H_\infty(c)$  such that no three of these vectors are collinear and sequences  $(\hat{h}_{T'}(c^{T'}))_{T' \geq T_C}$  such that for every  $c \in C$  and  $T \geq T_C$ ,  $T' \neq T$ ,  $\hat{h}_{T'}(c^{T'}) \in H_{T'}(c)^{T'}$  and  $\lim_{T' \rightarrow \infty} \hat{h}_{T'}(c^{T'}) = \hat{h}_\infty(c)$  (this can be done by Lemma A.8). Set  $h_{T'}(c^{T'}) = \hat{h}_{T'}(c^{T'})$  for all  $T' \geq T_C$ ,  $T' \neq T$ , and for any  $D' \in \mathbb{D}_{T'}$ , construct the vectors  $h_{T'}(D')$  as in (A-4).

The same procedure can be repeated for any  $h_T(D) \in H_T(D)$ , giving us a collection of functions  $\mathcal{H}_D$  for a specific database  $D \in \mathbb{D}^T$ . Note that the same sequences  $\left(\hat{h}_{T'}(c^{T'})\right)_{T \geq T_C}$  are used in the construction of each of these functions. The union of the sets  $\mathcal{H}_D$  over all databases in  $\mathbb{D}$ , gives us  $\mathcal{H} = \cup_{D \in \mathbb{D}} \mathcal{H}_D$ . It is obvious that these functions satisfy all of the conditions listed above.

Take any two such functions,  $h^1$  and  $h^2 \in \mathcal{H}$ . Suppose that  $h^1$  was constructed starting from a database  $D_1 \in \mathbb{D}_{T^1}$ , while  $h^2$  was constructed starting from a database  $D_2 \in \mathbb{D}_{T^2}$ . Then,  $h_{T'}^1 = h_{T'}^2$  for all  $T' \geq \max\{T^1; T^2; T_C\}$ .

From Proposition A.2, we conclude, that for each of the functions  $h \in \mathcal{H}$ , we can find a similarity function  $s^h$  and probability vectors  $\hat{p}_T^{h,c} = h_T(c^T)$  such that

$$h_T(D) = \frac{\sum_{c \in C} s^h(c) \hat{p}_T^{h,c} f_D(c)}{\sum_{c \in C} s(c) f_D(c)}$$

for all  $T \geq 2$  and all  $D \in \mathbb{D}^T$ . We set  $\hat{P}_T^c = \left\{ \hat{p}_T^{h,c} \right\}_{h \in \mathcal{H}} = \left\{ h_T(c^T) \right\}_{h \in \mathcal{H}} = H_T(c^T)$ .

We now show that  $s^h$  does not depend on the specific choice of the function  $h$ . Take any two functions  $h^1$  and  $h^2 \in \mathcal{H}$ . Let  $\bar{T}^1$  be the minimal value of  $T$  such that no three of the vectors  $(h_T^1(c^T))_{c \in C}$  are collinear for  $T \geq \bar{T}^1$ . From the proof of Lemma A.3, we know that  $\bar{T}^1$  is finite. Define  $\bar{T}^2$  analogously. Recall that the similarity functions  $s^1$  and  $s^2$  are defined by

$$h_{3\bar{T}^1}^1(f^{\{j,k,l\}}) = \frac{\sum_{m \in \{j,k,l\}} s_m^{1,\{j,k,l\}} \hat{p}_{3\bar{T}^1}^{1,m}}{\sum_{m \in \{j,k,l\}} s_m^{1,\{j,k,l\}}} \quad (\text{A-5})$$

$$h_{3\bar{T}^2}^2(f^{\{j,k,l\}}) = \frac{\sum_{m \in \{j,k,l\}} s_m^{2,\{j,k,l\}} \hat{p}_{3\bar{T}^2}^{2,m}}{\sum_{m \in \{j,k,l\}} s_m^{2,\{j,k,l\}}} \quad (\text{A-6})$$

where  $j, k$  and  $l$  is any triplet of cases in  $C$ . Since for all  $T' \geq \max\{T^1; T^2; T_C\}$ ,  $h_{T'}^1 = h_{T'}^2$ , there exists a  $k$  such that  $3k\bar{T}^1\bar{T}^2 \geq \max\{T^1; T^2; T_C\}$  and, hence, the equations

$$h_{3k\bar{T}^1\bar{T}^2}^1(f^{\{j,k,l\}}) = \frac{\sum_{m \in \{j,k,l\}} s_m^{1,\{j,k,l\}} \hat{p}_{3k\bar{T}^1\bar{T}^2}^{1,m}}{\sum_{m \in \{j,k,l\}} s_m^{1,\{j,k,l\}}} = \frac{\sum_{m \in \{j,k,l\}} s_m^{1,\{j,k,l\}} h_{3k\bar{T}^1\bar{T}^2}^1(c_m^{3k\bar{T}^1\bar{T}^2})}{\sum_{m \in \{j,k,l\}} s_m^{1,\{j,k,l\}}} \quad (\text{A-7})$$

$$h_{3k\bar{T}^1\bar{T}^2}^2(f^{\{j,k,l\}}) = \frac{\sum_{m \in \{j,k,l\}} s_m^{2,\{j,k,l\}} \hat{p}_{3k\bar{T}^1\bar{T}^2}^{2,m}}{\sum_{m \in \{j,k,l\}} s_m^{2,\{j,k,l\}}} = \frac{\sum_{m \in \{j,k,l\}} s_m^{2,\{j,k,l\}} h_{3k\bar{T}^1\bar{T}^2}^2(c_m^{3k\bar{T}^1\bar{T}^2})}{\sum_{m \in \{j,k,l\}} s_m^{2,\{j,k,l\}}} \quad (\text{A-8})$$

are equivalent. By Lemma A.4, the similarity values in equations (A-5) and (A-7) are identical, and so are the similarity values in equations (A-6) and (A-8). Hence, both  $h^1$  and  $h^2$  give rise to identical similarity functions.  $s^h$  is therefore independent of  $h$ .

We conclude that

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

■

## References

- AHN, D. (2008): “Ambiguity Without a State Space,” *Review of Economic Studies*, 71, 3–28.
- BEWLEY, T. F. (1986): “Knightian Decision Theory: Part 1,” Discussion Paper 807, Cowles Foundation.
- BILLOT, A., I. GILBOA, D. SAMET, AND D. SCHMEIDLER (2005): “Probabilities as Similarity-Weighted Frequencies,” *Econometrica*, 73, 1125–1136.
- BLIZARD, W. D. (1988): “Multiset Theory,” *Notre Dame Journal of Formal Logic*, 30(1), 36–66.
- CHATEAUNEUF, A., J. EICHBERGER, AND S. GRANT (2007): “Choice Under Uncertainty with the Best and the Worst in Mind: Neo-Additive Capacities,” *Journal of Economic Theory*, 137, 538–567.
- COIGNARD, Y., AND J.-Y. JAFFRAY (1994): “Direct Decision Making,” in *Decision Theory and Decision Analysis: Trends and Challenges*, ed. by S. Rios, Boston. Kluwer Academic Publishers.
- EICHBERGER, J., AND A. GUERDJIKOVA (2008): “Case-Based Expected Utility: Preferences over Actions and Data,” Discussion paper, Cornell University.
- ELLSBERG, D. (1961): “Risk, Ambiguity, and the Savage Axioms,” *Quarterly Journal of Economics*, 75, 643–669.
- EPSTEIN, L., AND M. SCHNEIDER (2007): “Learning Under Ambiguity,” *Review of Economic Studies*, 74, 1275–1303.
- GAJDOS, T., T. HAYASHI, J.-M. TALLON, AND J.-C. VERGNAUD (2007): “Attitude Towards Imprecise Information,” *Journal of Economic Theory*, 140(1), 27–65.
- GHIRARDATO, P., F. MACCHERONI, AND M. MARINACCI (2004): “Differentiating Ambiguity and Ambiguity Attitude,” *Journal of Economic Theory*, 118, 133–173.
- GILBOA, I. (2009): *Theory of Decision under Uncertainty*, Econometric Society Monographs. Cambridge University Press, Cambridge.
- GILBOA, I., O. LIEBERMAN, AND D. SCHMEIDLER (2006): “Empirical Similarity,” *Review of Economics and Statistics*, 88, 433–444.
- GILBOA, I., AND D. SCHMEIDLER (1989): “Maxmin Expected Utility with a Non-Unique Prior,” *Journal of Mathematical Economics*, 18, 141–153.
- GILBOA, I., AND D. SCHMEIDLER (2001): *A Theory of Case-Based Decisions*. Cambridge University Press, Cambridge, UK.
- GILBOA, I., D. SCHMEIDLER, AND P. WAKKER (2002): “Utility in Case-Based Decision Theory,”

- Journal of Economic Theory*, 105, 483–502.
- GONZALES, C., AND J.-Y. JAFFRAY (1998): “Imprecise Sampling and Direct Decision Making,” *Annals of Operations Research*, 80, 207–235.
- HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2001): *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Series in Statistics. Springer, New York.
- KEYNES, J. M. (1921): *A Treatise on Probability*. Macmillan, London.
- KLIBANOFF, P., M. MARINACCI, AND S. MUKERJI (2005): “A Smooth Model of Decision Making Under Ambiguity,” *Econometrica*, 73(6), 1849–1892.
- MANSKI, C. (2000): “Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice,” *Journal of Econometrics*, 95, 415–432.
- MANSKI, C. (2009): “Diversified Treatment under Ambiguity,” *International Economic Review*, p. forthcoming.
- MARINACCI, M. (2002): “Learning about Ambiguous Urns,” *Statistical Papers*, 43, 143–151.
- O’HAGAN, A., AND B. R. LUCE (2003): “A Primer on Bayesian Statistics in Health Economics and Operations Research,” Centre for Bayesian Statistics in Health Economics.
- PESKI, M. (2007): “Learning through Patterns,” Discussion paper, University of Chicago.
- ROCKAFELLAR, R. T., AND R. WETS (2004): *Variational Analysis*. Springer, Heidelberg.
- SCHMEIDLER, D. (1989): “Subjective Probability and Expected Utility without Additivity,” *Econometrica*, 57, 571–587.
- STINCHCOMBE, M. (2003): “Choice and Games with Ambiguity as Sets of Probabilities,” Discussion paper, University of Texas, Austin.