From Cases to Probabilities — Learning from Data with the Best and the Worst in Mind[1]

Jürgen Eichberger[2] and Ani Guerdjikova[3]

This version: March 8, 2010

Abstract

We consider a decision-situation characterized by a decision problem, i.e. a set of actions and a set of feasible outcomes and by an information context, i.e. a data-set containing the past performance of (some of) the actions. The decision-maker can express preferences over the choice of an action in a given context. Information contexts which contain a large number of observations and ing observations which are more relevant to the choice of the action under consideration are considered less ambiguous. We derive a representation of preferences, which separates utility and beiefs. While the utility function is fully subjective, the beliefs of the decision maker combine the objective characteristics of the data (number and frequency of observations) with the subjectively perceived relevance of observations (similarity). We identify the subjectively perceived degree of ambiguity and separate it into two components: the degree of optimism (the weight put on the best possible outcome), and the degree of pessimism (the weight assigned to the worst possible outcome). In a special case of our representation beliefs can be represented as similarity-weighted frequencies, thus providing a behavioral foundation of BGSS (2005).

[1] We would like to thank to Larry Blume, David Easley, Gaby Gayer, Itzhak Gilboa, John Hey, Edi Karni, Nick Kiefer, Peter Klibanoff, Mark Machina, Francesca Molinari, Ben Polak, Karl Schlag, David Schmeidler, conference participants at RUD 2008 in Oxford, FUR 2008 in Barcelona and at the Workshop on Learning and Similarity 2008 in Alicante, as well as seminar participants at Cornell for helpful suggestions and comments.

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

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

1 Introduction

Most of modern decision theory under uncertainty has been conducted in the framework of Savage (1954), who takes acts which associate outcomes with states as primitives and assumes that a decision maker can order these acts. Given his well-known axioms these preferences can be represented by a subjective expected utility functional, where the subjective probability distribution over states as well as the utility evaluation of outcomes are endogenously derived.

More recently, Gilboa and Schmeidler (2001) developed an alternative framework, case-based decision theory, in which actions and data-sets of cases are the primitive concepts. Preferences are assumed to be about actions conditional on information given in the data-set. Gilboa and Schmeidler (2001) also provide axioms which allow us to value an action by the frequency-weighted sum of its case-by-case evaluations.

Both approaches are behavioral, taking the observable preferences of the decision maker as primitive concept. They differ however in the objects which the decision maker is assumed to rank. For Savage (1954) all the information necessary for the evaluation of an action is encoded in the states. For Gilboa and Schmeidler (2001) information comes in the form of cases.

Gilboa and Schmeidler (2001) do not require the decision maker to condense the information from a data-set of cases into states of the world. This is an attractive feature of their approach. Their functional representation of preferences in the case-based context lacks however the intuitive appeal of the subjective expected utility approach of Savage (1954). In particular, it does not provide a clear separation of utility and beliefs.

Most recently, Billot, Gilboa, Samet and Schmeidler (2005), henceforth BGSS (2005), showed that one can derive probability distributions over outcomes as similarity-weighted frequencies of the observed cases. Eichberger and Guerdjikova (2009), henceforth EG (2009), extend this idea to a context of multiple-priors, hence allowing beliefs to also reflect the confidence of the decision maker in the informational content of the data. These results assume, however, a mapping from cases to probabilities over outcomes as primitive concepts and, therefore, lack a

behavioral foundation.

In the light of this literature, we suggest a behavioral approach which allows us to derive a representation of the subjective expected utility type. We deduce the mapping from data-sets to probability distributions over outcomes, which are exogenous concepts in BGSS (2005) and EG (2009), endogenously and apply it to an expected utility evaluation of actions. Furthermore, we relate the subjective beliefs of the decision maker to the objectively given information, by representing the probability distributions over outcomes as similarity-weighted frequencies of observations contained in the data-set. The key difference between the approach suggested in this paper and Gilboa and Schmeidler (2001) concerns the domain of preferences. Gilboa and Schmeidler (2001) assume a family of preference orders over actions. In our approach preferences are defined over the product space of actions and data-sets. In particular, we assume that the decision maker is able to make the following two types of comparisons: first, for a given data-set, he can express preferences for one course of action over another; second, for a given action, he can compare different data-sets based on the support they give to the choice of this particular action. We elaborate on this below.

The basic concepts of our approach are a *decision problem* and an *information context.* A *decision problem* (A, R) is characterized by a set of actions A and a set of possible outcomes R. An *information context* is uniquely defined by a data-set D of the form $D = (c_1 ... c_T) = ((a_1; r_1); ... (a_T; r_T))$, which represents the objective (i.e., interpersonally verifiable) information available to the decision maker in a given context[4]. The elements of the data-set are called cases $c = (a, r)$. Each case contains an action-outcome pair from the set of possible cases $C := A \times R$. The list of cases $(c_1 ... c_T)$ in a data-set D may contain records of action-outcome pairs observed in the past or cases derived from other sources. A data-set may be empty if the information context is completely novel. The set of all data-sets D for a given decision problem (A, R) is denoted by $\mathbb{D}$. A *decision situation* is completely described by the triple $(A, R, \mathbb{D})$,

[4] In practice, objective information will often be aggregated or condensed in descriptive statistics, indicators, or even vague narratives of past observations. This is an interesting avenue to pursue in further studies. At this stage, we abstract from such aggregation processes and restrict attention to the special case where the information context is exclusively described by a data set.

i.e., a decision problem and the set of possible information contexts arising from it.

Example 1. 1 Betting on a draw from an urn

A lottery offers bets on the color of the ball drawn from an urn containing black and white balls. There are two possible bets: a bet on a white ball, a_w *and a bet on a black ball,* a_b*,* $A = \{a_w; a_b\}$*. The outcome is a monetary prize of* 1 *if the color of the ball drawn corresponds to the bet, i.e.,* $r \in R = \{1, 0\}$*. An information context specifies the available information about an urn. It consists of a data-set* $D = ((a_1; r_1) ; ... (a_T; r_T))$ *containing records of bets and the outcomes of past drawings from the urn*[5]*.*

For instance, an urn could be described by the following data-set:

$$\begin{aligned} D_1 \quad = \quad & ((a_w; 1) ; (a_w; 1) ; (a_b; 0) ; (a_w; 0) ; (a_w; 1) ; \\ & (a_w; 0) ; (a_w; 0) ; (a_b; 1) ; (a_b; 1) ; (a_w; 1)) . \end{aligned} \tag{1}$$

Or, if the order of the draws does not matter, the data set could be summarized in the following table:

$$A \quad \begin{array}{c|cc} & R & \\ & 1 & 0 \\ \hline a_b & 2 & 1 \\ a_w & 4 & 3 \end{array} \tag{2}$$

■

Example 2. 1 Loan market

Consider an economy, in which entrepreneurs consider investing in risky projects, such as starting an internet retail company, opening a fast-food restaurant, etc. The set of such projects will be denoted by $a \in A$ *and their possible financial returns by* $r \in R$*. Entrepreneurs do not possess capital, but can borrow from creditors. In order to start a project, an entrepreneur needs one unit of capital.*

Each creditor has exactly one unit of capital. A standard credit contract specifies a fixed repayment q*, which is due whenever the payoff of the project exceeds it. Otherwise, the lender receives the entire return of the project.*

[5] The usual description of a statistical urn just specifies the number of balls of different colors contained in the urn. The information about the number of colors of otherwise identical balls summarizes the "ideal" environment for the aggregation of data sets of past observations by the frequencies of the balls in the urn.

Neither lenders nor entrepreneurs know the probability distribution over returns for a given project. Yet, both have access to past data about projects and their returns. The information about returns of projects for a specific market is described by a sequence of observations, e.g.,

$$D_1 = ((a_1; r_1) ... (a_T; r_T)) .$$

Two markets, which are similar except for the information available, may be distinguished by their information contexts reflected in different data sets, D_1, $D_2 \in \mathbb{D}$. For instance, D_1 may contain data from a well established market (e.g., a Western economy), while D_2 might contain the evidence from an emerging market (e.g., an Eastern European country).
It is possible that the empirical distributions of the returns of a project are very similar in both markets, yet the data-set for the emerging market contains fewer observations than the one for the well-established economy. The different quality of information may lead both investors and creditors to prefer the investments in the market with the more precise information. ■

Example 3. 1 Financial investment

Consider an investor who can invest 1 unit of money into one of several companies $a \in A$. Each company is described by the industry branch in which it operates, as well as by whether its assets are listed in the stock exchange. For instance, a_1 may stand for a large internet retailer listed in the stock exchange, while a_2 may represent an internet retailer, which is not listed, a_3 may stand for a non-listed food-store chain, etc. R represents the set of possible returns. The investor has access to past data containing information about the returns of the different companies. A data-set will describe the observed returns in a specific market. For instance, if $A = \{a_1; a_2\}$, the investor might have the following information about the returns of the last T periods for the market in his home-country:

$$D_H = \left(\left(a_1; r_1^1\right) ... \left(a_1; r_T^1\right); \left(a_2; r_1^2\right) ... \left(a_2; r_T^2\right)\right) \tag{3}$$

Different data-sets will describe asset performance in different markets. For instance, the investor might also have access to a second data-set D_F documenting the returns of companies of type a_1 and a_2 in a foreign market. The content of the two data-sets will differ depending

on the performance of companies in the two markets, but also on the ability of the investor to access information about these markets. If stock-market data are readily available for any country, while data on the performance of non-listed foreign companies are difficult to obtain, the investor might only have information about the listed company a_1, but not about a_2, in the foreign market F. Hence, the data-set for the foreign market for the same T periods would be given by the sequence of observations:

$$D_F = \left(\left(a_1; \tilde{r}_1^1\right) ... \left(a_1; \tilde{r}_T^1\right)\right). \tag{4}$$

■

Example 4. 1 Medical treatment

Consider a doctor who has to choose a treatment for a patient with a particular disease. The possible treatment options are administering a new drug, a_1, using the traditional treatment, a_2, or applying a placebo, a_3; hence, $A = \{a_1; a_2; a_3\}$. The potential outcomes reach from complete recovery, r_1, to several weeks of illness, r_2, or long-term chronic disease, r_3; i.e., $R = \{r_1; r_2; r_3\}$.

The information context captures the doctor's personal experience, results of clinical studies or records from hospitals. It can be represented by a data-set D consisting of cases $(a_t; r_t)$ which describe treatments chosen in the past and their outcomes. The following table shows a summary of a particular data-set D by listing the number of occurrences for each case:

		R		
		r_1	r_2	r_3
A	a_1	*15*	*20*	*0*
	a_2	*35*	*80*	*10*
	a_3	*5*	*70*	*15*

The data in this table reflect a limited experience with the new drug a_1 as compared to the greater experience with the traditional treatment a_2 and with no treatment a_3.

The doctor may also consider hypothetical data-sets which may result from new studies and tests. Such hypothetical information contexts could differ with respect to the quality and the type of observations, as well as with respect to the outcomes observed. ■

Decision makers compare actions belonging to different information contexts. Given a decision

situation $(A, R, \mathbb{D})$, they are supposed to be able to rank actions $a, a' \in A$ in their respective information contexts $D, D' \in \mathbb{D}$. Hence, $A \times \mathbb{D}$ is the domain of a decision maker's preference order $\succcurlyeq$. This implies, in particular, that a decision maker can express preferences over different data-sets for a given action $a \in A$. Ranking the action a higher in the information context D than in the information context D', $(a, D) \succcurlyeq (a, D')$, means that the decision maker would prefer the information context D over the information context D' when choosing action a. Such a preference reflects the common statement that, for choosing a particular action, "one would like to have more or better information". As the following examples show, it will depend on the decision situation whether "having more or better information" is a feasible option.

Example 1. 2 Betting on a draw from an urn (Example 1.1 continued)

In Ellsberg's two-urn experiment, a decision maker may have differing information about two urns which are ex-ante identical, i.e. contain the same total number of black and white balls. These urns can be represented by different information contexts — different data-sets D_1, $D_2 \in \mathbb{D}$. For instance, D_1 might be given by (2) which contains 10 observations, while there may be 300 observations in D_2, which are summarized in the following table:

		R	
		1	0
A	a_b	*80*	*90*
	a_w	*60*	*70*

(5)

The decision maker decides which urn to bet on and which bet to place. Hence, his preferences $\succsim$ are defined on $A \times \mathbb{D}$. For example, the decision maker may express preferences over the urn, i.e., the data-set D, from which the ball is drawn, as in Ellsberg's two-urn paradoox, $(a_b, D_2) \succ (a_b, D_1)$ and $(a_w, D_2) \succ (a_w, D_1)$.

Note that if two urns are described by the same data-set, e.g., D_2, then these urns are indistinguishable and represent the same information context. Hence, for any given bet, the decision maker should be indifferent between the two urns. In contrast, when the two urns are described by distinct information contexts, $D_1 \neq D_2$, a decision maker may well prefer the information context D_2 over the information context D_1 for all bets a in A. We conjecture that, as in the Ellsberg two-urn paradox, such behavior will be observed, whenever the data-set D_2 contains more precise information than D_1. ■

Example 2. 2 Loan market (Example 2.1 continued)

Consider two markets with identical (A, R) *but differing information* D_1 *and* D_2*. Entrepreneurs and lenders must agree on the type of project they want to undertake,* $a \in A$*, and in which of the markets they want to be active. Hence, they must be able to express preferences of the type* $(a_1; D_1) \succsim (a_2; D_2)$ *for projects* a_1 *and* a_2 *under the information contexts* D_1 *and* D_2*, respectively.* ■

Example 3. 2 Financial investment (Example 3.1 contd.)

Suppose that the investor has information about his home market and about a foreign market given by data-sets D_H *and* D_F *as in* (3) *and* (4)*. The investor could compare the prospects of investing in a listed internet retail company* a_1 *in his home market to investing in a similar company in a foreign market, expressing preferences of the type* $(a_1; D_H) \succsim (a_1; D_F)$*. This comparison will reflect the different quality of information about the two markets. In particular, if the investor considers the information about the performance of* a_2 *to be relevant for his prediction about the returns of* a_1*, he might feel that the data-set* D_H *is more informative than* D_F*. This might lead him to systematically prefer investments in his home market, giving rise to the well-known home-bias phenomenon.* ■

Example 4. 2 Medical treatment (Example 4.1 continued)

When choosing a treatment in A *for a given patient, the doctor must base her decision on the given data-set* D*, which defines the information context. Her decision will reflect her preferences over treatments* $(a_1; D)$*,* $(a_2; D)$ *and* $(a_3; D)$ *in the light of the data-set* D*, e.g.,*

$$(a_1; D) \succ (a_2; D) \succsim (a_3; D)\,.$$

Given the small sample of cases containing observations of the new drug a_1*, the doctor might also express preferences for obtaining additional data, for instance by conducting an additional study or buying another data-set* D'*:*

$$(a_1; D') \succsim (a_1; D)\,.$$

Such preferences for data-sets related to a given action a_1 *do not imply the availability of both*

data-sets. They are a hypothetical, but testable binary relation on data-sets D and D'. For instance, the evaluation of a potential study of 100 additional cases may require comparing the potential data-bases which could arise from such a study. ■

An interesting by-product of our approach is that it allows us to discuss preferences between different contexts, while keeping the action constant: $(a; D) \succsim (a; D')$. In the first three examples, preferences between different contexts correspond to observable choices, i.e. a choice of an urn, on which to place a bet, a choice of a country or a market, in which to invest. In these cases it appears reasonable to assume that the two information contexts are independent, i.e. learning about draws from urn I does not provide additional information about the composition of urn II. Similarly, learning about the performance of a certain type of investment in one country does not contribute to the evaluation of the investments available in a different country[6]. In the last example, however, the doctor's preferences between data-sets must be given a different interpretation — they reflect the difference in the degree of support of the choice of a provided by different information contexts. For instance, the doctor might state that she would feel more confident choosing a if a had performed better in those cases in which it was observed, or if the data-set contained more cases similar to a or if the data-set contained more observations. While the first statement is purely hypothetical, the second and the third do contain an element of choice. In fact, the doctor might suggest that a new clinical study be conducted so as to increase the number and relevance of observations in the already available data. Obviously, similar comparisons can be expressed for the other two examples, as well.

The examples, thus show that our framework admits for two modelling strategies concerning preferences for information: in the first case, the information context is a choice object itself, i.e., the information is identified with a physical choice entity (country, urn, etc.); in the second case, the decision maker hypothesizes about different information contexts, only some of which might actually be feasible. The first approach is behavioral, in that choices will reveal preferences across information contexts. The second approach relies to a large extent on cognitive

[6] The empirical fact that cross-country asset returns are uncorrelated is the main justification for the advantages of internationally diversified portfolios.

introspection, since the set of observable choices is much smaller. The choice of the modelling strategy will depend largely on the specific application and research question.

In view of the examples provided above, we consider a decision maker who has preferences defined on action-data-set-pairs. We provide necessary and sufficient conditions on this preference relation which ensure that the evaluation of action a given the information contained in data-set D can be written as an expected utility with respect to a von-Neumann-Morgenstern utility function over outcomes and a probability distribution over the outcomes of action a, conditional on the information context D. Thus, we separate the utility assigned to outcomes from the decision maker's beliefs. Furthermore, in the spirit of BGSS (2005), beliefs can be represented as a combination of the objective properties of the information context, such as empirical frequency and number of observations, and the subjective characteristics of the decision maker, such as his perceived degree of ambiguity, his attitude towards ambiguity and his similarity perceptions.

Our framework allows us to vary the degree of precision of the information context by increasing the length of the data-set, while keeping the empirical frequencies of observations unchanged. While such a change in the information precision is interpersonally verifiable, the perceived change in ambiguity will depend on subjective characteristics of the decision maker. The natural assumption that more precise data-sets are considered less ambiguous allows us to uniquely identify the decision maker's perceived degree of ambiguity as a function of the length of the data-set.

The decision maker's attitude towards ambiguity in the context of differing information precision can be described using Grant, Kaji and Polak's (1998, p. 234) quotation of the *New York Times*:

> "there are basically two types of people. There are "want-to-knowers" and there are "avoiders." There are some people who, even in the absence of being able to alter outcomes, find information of this sort beneficial. The more they know, the more their anxiety level goes down. But there are others who cope by avoiding, who would rather stay hopeful and optimistic and not have the unanswered questions answered."

By examining the decision maker's preferences over information contexts with differing degrees

of precision, we are able to uniquely identify the decision maker's attitude towards ambiguity — his degrees of optimism and pessimism in the spirit of the quotation above.

The decision maker's perception of similarity determines the relative weight assigned to different observations in the formation of beliefs. When the decision maker considers a data-set, in which distinct actions have been observed (as for instance in Example 4.1), it is natural to assume that cases containing the same action as the one under consideration will receive a higher weight in the evaluation than cases containing different actions. Identifying the similarity between two actions is based on comparisons between information contexts in which only one of the two actions has been observed to contexts in which both actions have been observed with positive frequency. Our axioms ensure that the similarity function can be uniquely determined from the decision maker's preferences.

The rest of the paper is organized as follows. The next section presents the framework and introduces the preference relation on action-data-set-pairs. Section 3 states the axioms. Section 4 derives the representation of preferences. In Section 5, we provide some examples which illustrate the representation. In Section 6, we relate our results to the existing literature. Section 7 concludes. All proofs are collected in the Appendix.

2 The Framework

Consider a decision problem $(A; R)$. The set of actions A and the set of outcomes are assumed to be finite with representative elements a and r, respectively. We deviate from the standard frameworks used in the literature to model decision-making under uncertainty. In particular, we assume that the decision maker knows neither the probability distribution of payoffs associated with a specific action a (as in the von-Neumann-Morgenstern model), nor the mapping which describes the state-contingent outcomes of an action as in Savage's framework. In contrast, all the information available to the decision maker is in form of observations. Each observation, i.e. *case*, consists of an action and an outcome generated from this action and we write:

$$c = (a; r)\,,\ a \in A,\ r \in R$$

for a specific case and

$$C = A \times R$$

for the set of all possible cases. A set of T such observations is referred to as a data-set of length T ($T \geq 1$):

$$D = (c_1 ... c_T) = ((a_1; r_1); ... (a_T; r_T)).$$

The set of all conceivable data-sets is denoted by $\mathbb{D}$. As discussed in the introduction, a data-set defines the information context of the decision problem $(A; R)$. The set of data-sets of length T is denoted by $\mathbb{D}^T$. The frequency of a data-set $D \in \mathbb{D}^T$ is given by:

$$f_D = (f_D(c))_{c \in C} = \left(\frac{|\{t \mid (a_t; r_t) = c\}|}{T} \right)_{c \in C}.$$

Note that we can associate a set of admissible frequencies F^T with each length T. These are the real-valued frequencies which can obtained from any combination of T cases:

$$f \in F^T =: \left\{ f \in \Delta^{|C|-1} \text{ with } f(c) \in \left\{ 0; \frac{1}{T} ... \frac{k}{T} ... \frac{T-1}{T}; 1 \right\} \text{ for all } c \in C \right\}.$$

$\delta_{(a';r')}$ denotes the frequency which puts the entire mass on the observation $(a'; r')$. For a given $\mu \in [0; 1]$, a convex combination of two frequencies $\mu f + (1 - \mu) f'$ is defined in the usual way. We will also use the notation $\mathbb{D}_a$ to denote the set of data-sets containing only observations of action a,

$$\mathbb{D}_a = \{D \in \mathbb{D} \mid f_D(a'; r) = 0 \text{ whenever } a' \neq a\}.$$

$\mathbb{D}_a^T$ stands for the set of data-sets of length T containing only observations of action a. The set of frequencies corresponding to the data-sets in $\mathbb{D}_a^T$ is then:

$$f \in F_a^T =: \left\{ f \in \Delta^{|R|-1} \text{ with } f(a; r) \in \left\{ 0; \frac{1}{T} ... \frac{k}{T} ... \frac{T-1}{T}; 1 \right\} \text{ for all } r \in R \right\}.$$

δ_r stands for the frequency which puts the entire mass on the outcome r.

We assume that the decision maker has preferences defined on the set $A \times \mathbb{D}$, denoted by $\succsim$.

3 Axioms

We now impose assumptions on preferences which will guarantee the desired representation. We first restrict the class of preferences to satisfy a non-triviality condition. This condition is

not necessary for the representation we wish to derive, however it ensures that certain elements of the representation can be uniquely determined.

Definition 3.1 $\succsim$ *is called non-trivial if*

(i) *either* $|R| = 3$*, and for any* $a \in A$ *and any two cases* c*,* $c' \in C$*, there exists a* $\bar{T} \in \mathbb{N}$ *and* D *and* D' *such that either*

$$\left(a; c^T\right) \succ (a; D) \succ (a; D') \succ \left(a; (c')^T\right),$$

or

$$\left(a; c'^T\right) \succ (a; D) \succ (a; D') \succ \left(a; c^T\right)$$

holds for any $T \geq \bar{T}$*.*

(ii) *or* $|R| > 3$ *and there is an* $a \in A$ *and two cases* c *and* $c' \in C$ *satisfying the condition in* (i)*.*

The definition restricts the class of preferences $\succsim$ by requiring that the decision maker distinguishes between evidence obtained from distinct cases as the number of observations becomes large. For the case of only three possible outcomes, we require that for sufficiently large T's, the T-fold observation of any two distinct cases provides distinct evidence w.r.t. to the choice of action a. Consider, e.g., two cases in which action a was observed, $c_1 = (a; r_1)$ and $c_2 = (a; r_2)$. If the decision maker were indifferent between obtaining r_1 and r_2, he might consider these two information contexts to be providing identical evidence w.r.t. a and express preferences

$$\left(a; c_1^T\right) \sim \left(a; c_2^T\right)$$

for all $T \in \mathbb{N}$. Hence, for $|R| = 3$, the Axiom precludes indifference among any two of the outcomes. It also excludes the case in which observing outcome r' from action a' is considered to provide identical information w.r.t. the outcome of a as observing outcome r'' from a different action a''. Hence, for the case of three outcomes, all cases have to convey distinct information with respect to any of the actions. For the case in which $|R| > 3$, the requirement is much weaker: there should be at least two cases which are considered to be distinct w.r.t. one of the alternatives. This assumption would be satisfied, if, e.g., at least two of the outcomes could be strictly ranked.

Axiom 1 (Complete order)

The preference relation $\succsim$ on $A \times \mathbb{D}$ is complete and transitive.

Axiom 1 is standard and without it a real-valued representation is impossible. While transitivity seems to be an innocuous assumption, completeness might be too demanding in this setting. In particular, it requires the decision maker to be able to imagine any two hypothetical data-sets D and D' and to be able to compare the prospects of any two actions a and a' with respect to these two data-sets.

Axiom 2 (Invariance)

Let π be a one-to-one mapping $\pi : \{1...T\} \to \{1...T\}$, then

$$\left(a; (c_t)_{t=1}^T\right) \sim \left(a; \left(c_{\pi(t)}\right)_{t=1}^T\right)$$

This is an exchangeability condition, which precludes a decision maker who is trying to discover serial correlation in the data. Axiom 2 implies that an information context is fully characterized by its length and the frequency of observations. For a data-set $D \in \mathbb{D}^T$ with frequency f_D, we can thus write

$$D = (f_D; T)\,.$$

From now on, we will use the notation D and $(f; T)$ interchangeably to denote an arbitrary data-set. As we noted above, however, not every combination $(f; T)$ defines a data-set. Hence, we have to require that $f \in F^T$. In what follows, whenever we state that a property holds for all $(f; T)$, we mean that it holds for all $(f; T)$ such that $f \in F^T$.

Axiom 3 (Betweenness for sets of equal length)

For any $a \in A$, T, $T' \in \mathbb{N}$, and $\mu \in (0; 1)$, if $(a; (f; T)) \overset{\succ}{(\sim)} (a; (f'; T))$, then

$$(a; (f; T')) \overset{\succ}{(\sim)} (a; (\mu f + (1-\mu) f'; T')) \overset{\succ}{(\sim)} (a; (f'; T'))\,.$$

To understand the Axiom, observe that for a given number of observations T, the decision maker prefers to choose a when the frequency in the data is f to the choice of a when the frequency is f'. Hence, his prediction about the outcome of a given f is more favorable than the prediction based on f'. The frequency of $\mu f + (1-\mu) f'$ is a linear combination of the two frequencies f and f', hence, it contains the favorable evidence from f and the less favorable from f' in some proportions μ and $(1-\mu)$. As long as the length of all three data-sets is equal, $\mu f + (1-\mu) f'$ should be ranked between f and f'. The Axiom suggests that the length of the data-set can be separated from the frequency of observations, when evaluating the information

context w.r.t. the choice of a. Intuitively, the frequency determines how much support the information context gives w.r.t. the choice of the action a, while the length determines the precision of the information. Keeping the precision constant across two information contexts, they can be ranked based solely on their content. Hence, the preference between $(a;(f;T))$ and $(a;(f';T))$ should not change, if the length of the two sets is changed to T', while keeping the frequencies unchanged. In contrast, changing the precision might lead to a reversal of preferences. For instance, a decision maker might have preferences $(a;(f';T'')) \succ (a;(f;T))$ for some $T'' > T$, if he thinks that the greater precision of $(f';T'')$ compensates for the weaker support this evidence provides for a.

Axiom 4 (Independence) For all T', $T'' \in \mathbb{N}$, all $\hat{a}'$ and $\hat{a}'' \in A$, all f_1', $f_2' \in F_{a'}^{T'}$, f_1'', $f_2'' \in F_{a''}^{T''}$ and $\mu \in [0;1]$,

$$(\hat{a}';(f_1';T')) \underset{(\prec)}{\succ} (\hat{a}'';(f_1'';T'')) \tag{6}$$
$$(\hat{a}';(f_2';T')) \underset{(\prec)}{\succsim} (\hat{a}'';(f_2'';T''))$$

implies:

$$(\hat{a}';(\mu f_1' + (1-\mu)\, f_2';T')) \underset{(\prec)}{\succ} (\hat{a}'';(\mu f_1'' + (1-\mu)\, f_2'';T'')) \tag{7}$$

and if $(\hat{a}';(f_2';T')) \sim (\hat{a}'';(f_2'';T''))$, then the two statements are equivalent.

This is an independence Axiom, which states that if the decision maker prefers the choice of $\hat{a}'$ given $(f_1';T')$ to the choice of $\hat{a}''$ given $(f_1'';T'')$, and, similarly, prefers $\hat{a}'$ given $(f_2';T')$ to the choice of $\hat{a}''$ given $(f_2'';T'')$, then the choice of $\hat{a}'$ given the linear combination of $(f_1';T')$ and $(f_2';T')$ is preferred to $\hat{a}''$ given the linear combination (with the same coefficient μ) of $(f_1'';T'')$ and $(f_2'';T'')$. The Axiom thus postulates that the weight given to the evidence in $(f_1';T')$ in the evaluation of $(\mu f_1' + (1-\mu)\, f_2';T')$ w.r.t. action $\hat{a}'$ is the same as the one of $(f_1'';T)$ in the evaluation of $(\mu f_1'' + (1-\mu)\, f_2'';T'')$ w.r.t. $\hat{a}''$. This appears intuitive, since the sets $(f_1';T')$ and $(f_2';T')$ are of the same length. Hence, the decision maker cannot distinguish between the two data-sets based on their precision. Furthermore, we assume that the decision maker will consider cases containing the observation of the same action to be equally relevant, regardless of the outcomes observed. This precludes behavior which is based, e.g., on consistently assigning higher weights to cases containing good outcomes. This implies that when combining the

frequencies of the two initial data-sets in proportion μ to $(1-\mu)$, the decision maker will put a weight of μ on the evidence from the first data-set and a weight of $(1-\mu)$ on the evidence from the second. Since the same argument applies to $(f_1''; T'')$ and $(f_2''; T'')$, the independence assumption appears reasonable.

Axiom 5 (Action-independent evaluation of payoffs) For all $T \in \mathbb{N}$, all a, $a' \in A$, $f \in F_a^T$, $f' \in F_{a'}^T$ such that $f(a;r) = f(a';r)$ for all $r \in R$,

$$(a;(f;T)) \sim (a';(f';T)).$$

Axiom 5 ensures that the decision maker chooses between actions based entirely on the available observations. Whenever two actions have performed identically for the same number of periods, their evaluation is the same. The Axiom, thus allows us to separate the evaluation of payoffs from the specific action they have been resulted from and guarantees that a utility function over outcomes can be derived.

Axiom 6 (Most favorable and least favorable outcome)

There exist $\bar{r}$ and $\underline{r} \in R$ such that for all $T \in \mathbb{N}$, all $a \in A$, and all $D \in \mathbb{D}^T$,

$$\left(a;(a;\bar{r})^T\right) \succsim (a;D) \succsim \left(a;(a;\underline{r})^T\right).$$

This Axiom postulates that there is an outcome $\bar{r}$ which, when observed T times in combination with a, provides the best possible evidence in favor of choosing a among all data-sets with length T. In other words, the decision maker will always prefer the choice of a in the information context $(a;\bar{r})^T$ to the choice of a given any other collection of T observations. Similarly, there is an outcome $\underline{r}$, which observed T times in combination with a provides the worst possible evidence for a among all data-sets of length T. Intuitively, observing the worst possible outcome in combination with the action under evaluation, a identifies the worst possible context, in which a might be chosen (as compared, e.g., to observing the performance of different actions a') and similarly, for the best possible outcome. By Axiom 5, the best (and the worst) outcomes will coincide for all actions $a \in A$, hence, we choose to disregard the potential dependence of $\bar{r}$ and $\underline{r}$ on a in the statement of the axiom.

Axiom 7 (Continuity) If

$$\left(a;(a;\bar{r})^T\right) \succ (a;D) \succ \left(a;(a;\underline{r})^T\right)$$

there is a $\kappa \in \mathbb{N}$ and μ, $\nu \in \left\{\frac{1}{\kappa T}; \frac{2}{\kappa T} ... \frac{\kappa T-1}{\kappa T}\right\}$ such that

$$\left(a; \left(\mu\delta_{(a;\bar{r})} + (1-\mu)\,\delta_{(a;\underline{r})}; \kappa T\right)\right) \succ (a; D) \succ \left(a; \left(\nu\delta_{(a;\bar{r})} + (1-\nu)\,\delta_{(a;\underline{r})}; \kappa T\right)\right).$$

The Axiom requires that the evaluation of $(a; D)$ can be approximated arbitrarily closely by data-sets containing only observations of a in combination with only the best and the worst outcome. It excludes the case in which the choice of a given information context D is worse than choosing a after T realizations of the best outcome, however, the observation of even one realization of the worst outcome in an arbitrarily large data-set, containing otherwise only realizations of the best outcome for a, would make the choice of a given D appear better.

Axiom 8 (Length-independence)

Suppose that for some $\bar{f}$, $\bar{f}'$ and $\bar{f}'' \in F_{\bar{T}}$ be such that $\left(a; \left(\bar{f}; \bar{T}\right)\right) \succsim \left(a; \left(\bar{f}''; \bar{T}\right)\right) \succsim \left(a; \left(\bar{f}'; \bar{T}\right)\right)$. Let f, g, f' and g', as well as $\hat{f}$, $\hat{g}$, $\hat{f}'$ and $\hat{g}'$ denote arbitrary frequencies in F_a^T.

If the statements

$$(a; (f; T)) \succsim \left(a; \left(\bar{f}; \bar{T}\right)\right) \succsim (a; (g; T))$$

and

$$(a; (f'; T)) \succsim \left(a; \left(\bar{f}'; \bar{T}\right)\right) \succsim (a; (g'; T))$$

imply that

$$(a; ([\lambda f + (1-\lambda)\, f']; T)) \succsim \left(a; \left(\bar{f}''; \bar{T}\right)\right) \succsim (a; (\lambda g + (1-\lambda)\, g'; T))$$

holds for some $\lambda \in (0; 1)$, then for any $\hat{T} \in \mathbb{N}$, the statements

$$\left(a; \left(\hat{f}; T'\right)\right) \succsim \left(a; \left(\bar{f}; \hat{T}\right)\right) \succsim (a; (\hat{g}; T'))$$

and

$$\left(a; \left(\hat{f}'; T'\right)\right) \succsim \left(a; \left(\bar{f}'; \hat{T}\right)\right) \succsim (a; (\hat{g}'; T'))$$

imply

$$\left(a; \left[\lambda\hat{f} + (1-\lambda)\,\hat{f}'\right]; T'\right) \succsim \left(a; \left(\bar{f}''; \hat{T}\right)\right) \succsim (a; (\lambda\hat{g} + (1-\lambda)\,\hat{g}'; T')).$$

The role of Axiom 8 is twofold. First, it ensures that the relevance of a case in D for the evaluation of an action a does not depend on the length of the data-set. Second, it guarantees that the evaluation of outcomes is independent of the length of the data-set. To illustrate these two features, we discuss two special cases of the Axiom.

First, assume that $\mu\bar{f} + (1-\mu)\,\bar{f}' = \bar{f}''$. Suppose also that for some f, $f' \in F_a^T$ and $\lambda \in (0; 1)$, we have

$$(a; (f; T)) \sim \left(a; \left(\bar{f}; \bar{T}\right)\right) \tag{8}$$

$$(a;(f';T)) \sim \left(a;\left(\bar{f}';\bar{T}\right)\right)$$

and

$$(a;([\lambda f + (1-\lambda) f'];T)) \sim \left(a;\left(\bar{f}'';\bar{T}\right)\right).$$

We have thus found data-sets which contain only observations of action a and which provide exactly the same information about the prospective payoffs of a as $\bar{f}$, $\bar{f}'$ and $\bar{f}''$. Furthermore, we know that when the data from $\left(\bar{f};\bar{T}\right)$ and $\left(\bar{f}';\bar{T}\right)$ are combined in proportions μ and $(1-\mu)$, the respective equivalent combination of f and f' has weights λ and $(1-\lambda)$. Since these two data-sets have equal length, the difference between the coefficients λ and μ can be only due to the fact that the relevance of information contained in f and f' is different. For instance, if $f = \delta_{(a;r)}$, whereas $f' = \delta_{(a';r)}$ for some $a' \neq a$, we would expect f to be more relevant for the evaluation of a than f' and, hence, $\lambda > \mu$ would obtain. The Axiom now requires the so found λ to be independent of the number of observations: if we have found the data-sets $\left(\bar{f};\hat{T}\right)$ and $\left(\bar{f}';\hat{T}\right)$ to be equivalent in terms of the information they provide for a to $\left(\hat{f};T'\right)$ and $\left(\hat{f}';T'\right)$, then the equivalent of $\left(\bar{f}'';\hat{T}\right)$ is the linear combination of these equivalents with the same factor λ:

$$\begin{aligned} \left(a;\left(\hat{f};T'\right)\right) &\sim \left(a;\left(\bar{f};\hat{T}\right)\right) \\ \left(a;\left(\hat{f}';T'\right)\right) &\sim \left(a;\left(\bar{f}';\hat{T}\right)\right) \end{aligned} \tag{9}$$

imply

$$\left(a;\left(\left[\lambda\hat{f} + (1-\lambda)\,\hat{f}'\right];T'\right)\right) \sim \left(a;\left(\bar{f}'';\hat{T}\right)\right).$$

Intuitively, the relevance of a case for the evaluation of an action is based on some a priori information, which is encoded in the structure of the action set A and which cannot be learned from observing the evidence in the data. Hence, it does not depend on the number of observations. While the change in the number of observations will certainly influence the evaluation of $\left(a;\left(\bar{f};\hat{T}\right)\right)$ and $\left(a;\left(\bar{f}';\hat{T}\right)\right)$, once these evaluations have been fixed, the evaluation of $\left(a;\left(\bar{f}'';\hat{T}\right)\right)$ can be determined as a weighted average of the two with the fixed coefficient λ found above.

In general, we are not able to find exact equivalents for the data-sets $\left(\bar{f};\bar{T}\right)$, $\left(\bar{f}';\bar{T}\right)$and $\left(\bar{f}'';\bar{T}\right)$,

since the set of data-sets is non-convex (it does not include irrational-valued frequencies). However, the Axiom allows us to replace the equivalents derived above and used in equations (8) and (9) with an approximation of the equivalents from above and from below. The same reasoning applies in this case and the Axiom ensures that the relevance of distinct cases remains constant as the number of observations changes.

The second role of the Axiom is to ensure that the evaluation of outcomes does not depend on the number of observations. To see this, suppose that $\bar{f} = \delta_{(a;\bar{r})}$, $\bar{f}' = \delta_{(a;\underline{r})}$, $\bar{f}'' = \delta_{(a;r)}$. If we can find a λ such that

$$\left(a; \lambda\delta_{(a;\bar{r})} + (1-\lambda)\,\delta_{(a;\underline{r})}; \bar{T}\right) \sim \left(a; \left(\delta_{(a;r)}; \bar{T}\right)\right), \tag{10}$$

then

$$\left(a; \lambda\delta_{(a;\bar{r})} + (1-\lambda)\,\delta_{(a;\underline{r})}; \hat{T}\right) \sim \left(a; \left(\delta_{(a;r)}; \hat{T}\right)\right)$$

holds for all $\hat{T}$. Intuitively, the data-sets $\bar{f}$, $\bar{f}'$ and $\bar{f}''$ differ only in the observed outcomes, but not in their length. Hence, the coefficient λ reflects the evaluation of the outcome r relative to the best and the worst outcome, $\bar{r}$ and $\underline{r}$. But the perceived utility of an outcome is a time-invariant characteristic of the decision maker which is independent of the length of the data-set. Hence, the coefficient λ should be independent of the number of observations.

Axiom 9 (Neutral outcome)

For all $a \in A$, there exists a $\hat{r} \in R$ such that for all $k > 1$,

$$(a; (a; \hat{r})) \sim \left(a; (a; \hat{r})^k\right).$$

This Axiom postulates the existence of an outcome, such that additional observations of this outcome in combination with action a do not provide additional evidence with respect to the performance of action a. Note that if we allowed for non-rational valued frequencies, Axioms 1-8 imply that there would always be a frequency f_a with $f_a \in F_a$ such that:

$$(a; (f_a; T)) \sim \left(a; (f_a; T)^k\right)$$

for any $k \in \mathbb{N}$. Intuitively, the decision maker regards any number of repetitions of the data-set $(f_a; T)$ as being equivalent to an objective randomizing device with a probability distribution over outcomes given by f_a. One could then augment the outcome space by an outcome $\hat{r}$

which would be the decision maker's certainty equivalent of f_a to guarantee that the statement of Axiom 9 holds. As before, by Axiom 5, $\hat{r}$ is identical for all actions a and we omit the dependence on a in the statement of the axiom.

Axiom 10 (Decreasing ambiguity)

For all $a \in A$, $T \in \mathbb{N}$ and all $D \in \mathbb{D}$ such that

$$\left(a; D^k\right) \succ \left(a; \left(a; \hat{r}_a\right)\right),$$

for some $k \in \mathbb{N}$, $\left(a; D^{T+1}\right) \succ \left(a; D^T\right)$.

For all $D \in \mathbb{D}$ such that

$$\left(a; \left(a; \hat{r}_a\right)\right) \succ \left(a; D^k\right),$$

for some $k \in \mathbb{N}$, $\left(a; D^T\right) \succ \left(a; D^{T+1}\right)$.

The last Axiom establishes the connection between the precision of a data-set and its content. If the choice of a given some data-set D is preferred to choosing a given an observation of the neutral outcome, then, intuitively, the content of D provides some positive evidence for choosing a. As the information of this data-set becomes more precise (the number of observations increases, while the frequency remains constant), the new data-set D^k will provide even stronger evidence in favor of a. In contrast, if the data-set under consideration represents evidence which is worse than the neutral outcome w.r.t. a, then as its precision increases (while the frequency remains constant), the evidence w.r.t. to a should become even more negative.

4 The Representation

Theorem 4.1 *A non-trivial preference relation $\succsim$ on $A \times \mathbb{D}$ satisfies Axioms 1–10 if and only if*

1. *there exists a maximal with respect to set inclusion correspondence $\hat{P} : N \times A \times C \rightrightarrows \Delta^{|R|-1}$, a similarity function $s : A \times A \to R^+ \setminus \{0\}$ and a utility function over outcomes $u : R \to R$, such that $\succsim$ can be represented by:*

$$V\left(a; D\right) = \sum_{r \in R} u\left(r\right) h(r|a, D), \tag{11}$$

$$h(r|a, D) := \sum_{c \in C} \frac{s_a\left(a_c\right) f_D\left(c\right)}{\sum_{c \in C} s_a\left(a_c\right) f_D\left(c\right)} \hat{p}^c_{T,a}\left(r\right) \quad \textit{for } r \in R. \tag{12}$$

Here, a_c is the action chosen in case c, $f_D\left(c\right)$ is the frequency with which case c is observed in D, T is the length of D, and $\hat{p}^c_{T,a}$ is any probability distribution over outcomes in the image of $\hat{P}^c_{T;a}$. The utility function u is unique up to an affine linear transformation, the sim-

ilarity function s_a is unique up to a multiplication by a positive number and the probability correspondence $\hat{P}$ is unique.

2. *Furthermore, for any (a, c), $\lim_{T\to\infty} \hat{P}^c_{T,a} = H^c_a$ exists and there are selections $\hat{p}^c_{T,a} \in \hat{P}^c_{T,a}$ such that, for all $c \in C$, $\lim_{T\to\infty} \hat{p}^c_{T,a} = h^c_a \in H^c_a$ exists and*

$$\hat{p}^c_{T,a} = \gamma_T \left(\alpha\delta_{\bar{r}} + \beta\delta_{\underline{r}}\right) + \left(1 - \gamma_T\right) h^c_a, \tag{13}$$

where $h^{(a;r)}_a(r) = \delta_r$ for all $r \in R$.
The parameters α, β and $(\gamma_T)^\infty_{T=1}$ are unique and satisfy α, β, $\gamma_T \in [0, 1]$ for all T and $\alpha+\beta = 1$. The sequence $\{\gamma_T\}^\infty_{T=1}$ is strictly monotonically decreasing and $\lim_{T\to\infty} \gamma_T = 0$.

The representation in Equations (11) and (12) can be interpreted in the following way:

(i) For an action $a \in A$ and a data set $D \in \mathbb{D}$, there is an expected-utility representation, Equation (11), where the probability distribution $h(r|a, D)$ is a selection from a set of probability distributions over outcomes for which the decision maker is indifferent.

(ii) The representation of beliefs in Equation (12) decomposes the beliefs of the decision maker into

- the frequencies of the cases in the data set $f_D(c)$,
- a function $s_a(a_c)$ reflecting the similarity between the act under consideration a and the act a_c in case c, and
- the predicted probability distribution over outcomes $\hat{p}^c_{T,a}(r)$ for the act under consideration a based on the special data set containing the T-fold observation of case c.

While the frequencies f_D represent the objective element of the prediction, both the similarity function s_a and the predicted probability distributions $\hat{p}^c_{T,a}$ are subjective elements of it. The similarity function can be viewed as an element of the belief about which one can argue based on (not explicitly modelled) aspects of the decision context such as company characteristics in an investment problem or similarities in the chemical composition of different drugs. The predicted probability distributions $\hat{p}^c_{T,a}$, however, reflects the irreducibly subjective assessment of the probability distribution over outcomes which the decision maker attributes to the act a if a data set consisting of a single case c observed T-times were given. If the action in case c, a_c, were equal to the act under consideration a, this would represent the ideal situation where the decision maker is faced with evidence from a direct test of the action a. In this case, the limiting distribution $h^c_a(r) = h^{(a;r)}_a(r)$ will predict outcome r for action a with probability 1,

δ_r. In general, however, the act a_c in case c will be different from the act a which is evaluated and, hence, the predicted probability distribution over outcomes h_a^c will not be concentrated on a particular outcome. $h_a^{(a';r')}(r)$ represents the decision maker's conditional assessment of the probability that the choice of a will result in outcome r conditional on the observation of $(a';r')$. In our framework, each case contains the outcome of a single action, hence, the correlation across actions is not observed in the data. $h_a^{(a';r')}$ is, thus, purely subjective. For the same reason, this is also the only element of our representation which is not uniquely identified: any element of the set of limit probability distributions $H_a^{(a';r')}$ will satisfy the properties of our representation.

Clearly, this decomposition of beliefs is hypothetical in the sense that the decision maker actually faces the data set D and not data sets c^T consisting of T-fold observations of a single case c. The decomposition clarifies, however, how the predicted probability distribution over outcomes $h(r|a,D)$ combines the objective properties of a data set D, such as frequencies and length, with subjective elements, such as similarity of acts, and personal characteristics, such as *ambiguity* and *ambiguity attitude*.

The dependence of $\hat{p}_{T,a}^c$ on T reflects the fact that longer data-sets are a more reliable source of information. In particular, the larger the number of observations of a given case, c, the more confident can the decision maker be about the outcome of action a_c. If $a_c = a$, then repeated observations of $(a;r)$ make the decision maker more confident that the probability distribution associated with a is concentrated on r. Equation (13) captures this idea by assigning a weight of $(1-\gamma_T)$ to the probability distribution h_a^c which the decision maker assigns to act a in the limit based on a sequence of observations c. The coefficient γ_T can be interpreted as the *perceived degree of ambiguity* for a given number of observations T. It decreases with T and eventually converges to 0. With an increasing number of observations of the case $(a;r)$, the decision maker becomes more and more convinced that the outcome of a will be the probability distribution h_a^c. On the other hand, if the number of observations is relatively small, the decision maker will assign some weight γ_T to a weighted average of the best and the worst possible outcomes, $\overline{r}, \underline{r}$,

reflecting a degree of *optimism* α and of *pessimism* β.

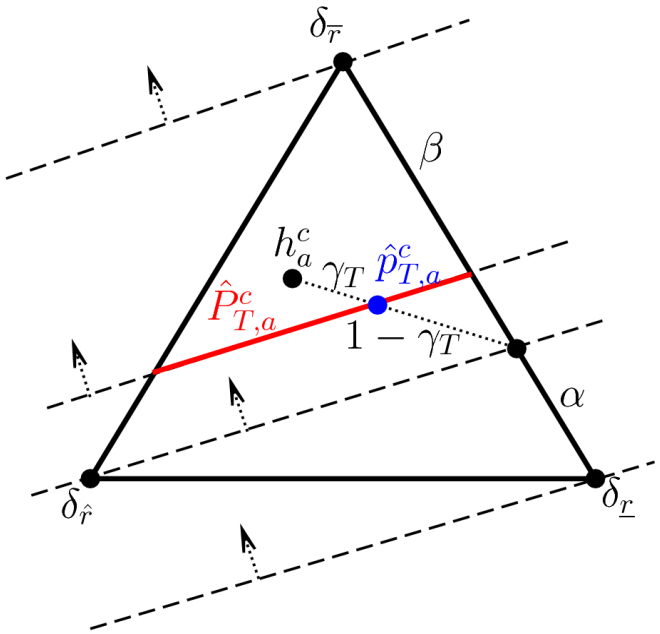

Figure 1. Probability Simplex

For the case of three outcomes, $R = \{\overline{r}, \underline{r}, \widehat{r}\}$, Figure 1 shows how the ambiguity attitudes of the decision maker are represented by the weights α and β given to the prediction of the best and worst outcomes, $\overline{r}, \underline{r}$, respectively. The set of probability distributions $\hat{P}_{T;a}^c$ for which a decision maker is indifferent given the data set D is represented in the diagram by the red line through the selected point $\hat{p}_{T,a}^c$. The neutral outcome $\hat{r}$ which was postulated in Axiom 9 determines the optimism-pessimism parameters α, β. Decreasing ambiguity with a growing number of observations T, which is postulated in Axiom 10, determines the parameter γ_T. Both axioms appear to be testable in behavioral experiments.

Remark 4.1 *In contrast to the literature in the Savage framework which, following* GILBOA & SCHMEIDLER (1989)*, models ambiguity by multiple priors, here ambiguity is associated with a particular subjective attitude towards the number of cases in a data set. Hence, one can relate the degree of ambiguity of a decision maker to the number of observations in a data set. Presenting subjects with data sets containing cases with equal frequencies but supported by differing lengths, one can hope to measure their degree of ambiguity. Hence, neutrality towards ambiguity means independence from the number of observations in data sets.*
In order to pin down the attitude towards ambiguity by parameters it is necessary to fix some reference point. In this paper, this is done by postulating a neutral outcome in Axiom 9. There

are other ways imaginable for determining ambiguity attitude. ■

We conclude this section with a variation on Axiom 9 which characterizes decision makers who are *ambiguity-neutral.* Ambiguity-neutral agents consider data-sets with equal frequencies to be equally informative without regard to their lengths.

Axiom 9A (All outcomes are neutral)

For all $a \in A$, and for all $\hat{r} \in R$,

$$(a;(a;\hat{r})) \sim \left(a;(a;\hat{r})^k\right)$$

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

Corollary 4.2 *$\succsim$ on $A \times \mathbb{D}$ satisfies Axioms 1–8 and 9A if and only if there exists a maximal with respect to set inclusion correspondence $\hat{P} : A \times C \rightrightarrows \Delta^{|R|-1}$, a similarity function $s : A \times A \to \mathbb{R}^+ \setminus \{0\}$ and a utility function over outcomes $u : R \to \mathbb{R}$, such that $\succsim$ can be represented by:*

$$V(a;D) = \sum_{r \in R} u(r)\, h(r|a,D), \tag{14}$$

$$h(r|a,D) : = \sum_{c \in C} \frac{s_a(a_c)\, f_D(c)}{\sum_{c \in C} s_a(a_c)\, f_D(c)} \hat{p}_a^c(r) \quad \text{for } r \in R, \tag{15}$$

where a_c is the action chosen in case c, $f_D(c)$ is the frequency with which case c is observed in D and $\hat{p}_a^c$ is any probability distribution over outcomes in the image of $\hat{P}_a^c$. The utility function u is unique up to an affine linear transformation, the similarity function s_a is unique up to a multiplication by a positive number and the probability correspondence $\hat{P}$ is unique. Furthermore, the sets $\hat{P}_a^{(a;r)}$ have the form

$$\hat{P}_a^{(a;r)} = \left\{h \in \Delta^{|R|-1} \mid h \cdot u = u(r)\right\}.$$

This representation associates identical probability distributions over outcomes with data-sets of *different length*, but *identical frequencies*. It captures the case of a decision maker who does not value the additional precision of a data-set with more observations. It is in this sense that the representation in Equation (14) provides a behavioral foundation for the result of BGSS (2005).

5 Examples with the representation

In this section we reconsider the examples introduced in Section 1 and show how the represen-

tation function in Equation (11) can be applied to these decision problems.

Example 1. 3 Betting on a draw from an urn (Example 1.1 continued)

*This example provides a generalization of Ellsberg's two-urn paradox. Instead of stating that "*there are 50 white and 50 black balls in Urn II*" and providing no information about Urn I, we provide data for both urns. The two data-sets are characterized by identical frequencies of black and white, however, the fact that there are more observations for Urn I means that the information for this urn is more precise than the information available for Urn II. Ellsberg's experiment requires subjects to compare bets on Urn I,* (a_b, D_1) *and* (a_w, D_1)*, with bets on Urn II,* (a_b, D_2) *and* (a_w, D_2)*.*

Suppose that the information about Urn I and Urn II is given by the data sets D_1 *and* D_2 *in Tables (*2*) and (*5*), respectively. The data set* D_2 *contains considerable evidence for frequencies of black and white balls being close to* $(\frac{1}{2}, \frac{1}{2})$ *in Urn II. Betting on black in Urn II yields a frequency of* $\frac{8}{17}$ *for black and a frequency of* $\frac{9}{17}$ *for white. Similarly, the data about bets on white reveal frequencies of* $\frac{6}{13}$ *and* $\frac{7}{13}$ *for black and white, respectively. Moreover, if decision makers aggregate cases by adding the frequency of bets on black lost,* $(a_b, 0)$*, and bets on white won,* $(a_w, 1)$*, since winning with white implies losing with black, they obtain a frequency of drawing a white ball from Urn II of* $\frac{1}{2}$*. Notice, however, that this aggregation relies on information about the decision context which is not encoded in the data set* D_2*, namely that there are only black and white balls in the urn.*

Information about Urn I is recorded in data set D_1*. Reasoning as before, one can deduce a frequency for black balls of* $\frac{2}{3}$ *from the bets on black and of* $\frac{3}{7}$ *from the bets on white. Aggregating as before, one finds again a frequency of* $\frac{1}{2}$ *for black balls. For Urn I, however, the frequencies of black and white draws are based on much less evidence (*10 *cases) than for Urn II (300 cases). The representation in Equation (*11*) yields*

$$V(a_b; D) = u(1)h(1|a_b, D) + u(0)h(0|a_b, D) \tag{16}$$

with

$$h(r|a_b, D) = \left(\frac{s_{a_b}(a_b)\left[f_D(a_b,1)\hat{p}_{T,a_b}^{(a_b,1)}(r) + f_D(a_b,0)\hat{p}_{T,a_b}^{(a_b,0)}(r)\right]}{s_{a_b}(a_b)\left[f_D(a_b,1)+f_D(a_b,0)\right] + s_{a_b}(a_w)\left[f_D(a_w,1)+f_D(a_w,0)\right]} + \frac{s_{a_b}(a_w)\left[f_D(a_w,1)\hat{p}_{T,a_b}^{(a_w,1)}(r) + f_D(a_w,0)\hat{p}_{T,a_b}^{(a_w,0)}(r)\right]}{s_{a_b}(a_b)\left[f_D(a_b,1)+f_D(a_b,0)\right] + s_{a_b}(a_w)\left[f_D(a_w,1)+f_D(a_w,0)\right]} \right).$$

In this situation, the decision maker is likely to consider the relevance of betting on white a_w, and betting on black a_{a_b}, to be equal for the prediction of the outcomes, i.e., $s_{a_b}(a_b) = s_{a_b}(a_w) = 1$. This assumption is suggestive in the context of betting on urns, where the events are comprehensive and mutually exclusive.

On the other hand, there is good reason for the decision maker to take into account the amount of evidence T available in the information contexts D_1 and D_2 of Urn I and Urn II. In this urn example, winning with a bet on black and losing with a bet on white, i.e., $(a_b, 1)$ and $(a_w, 0)$, yield the same information about the probability of winning with a bet on black a_b. Hence, one has

$$h_{a_b}^c = \begin{cases} \delta_1 & \text{for } c = (a_b,1) \text{ or } \quad c = (a_w,0) \\ \delta_0 & \text{for } c = (a_b,0) \text{ or } \quad c = (a_w,1) \end{cases}$$

and from Equation (13), $\hat{p}_{T,a}^c(r) = \gamma_T\left(\alpha\delta_{\bar{r}}(r) + \beta\delta_{\underline{r}}(r)\right) + (1-\gamma_T)\, h_a^c(r)$, one gets:

$$T = \begin{cases} |D_1| = 10 \\ |D_1| = 300 \end{cases} :$$

	$\hat{p}_{T,a}^c(r)$	
$\mathbf{a} = \mathbf{a}_b$	$\mathbf{r} = \mathbf{1}$	$\mathbf{r} = \mathbf{0}$
$\mathbf{c} = (\mathbf{a}_b, \mathbf{1})$	$\alpha\gamma_T + (1-\gamma_T)$	$\beta\gamma_T$
$\mathbf{c} = (\mathbf{a}_b, \mathbf{0})$	$\alpha\gamma_T$	$\beta\gamma_T + (1-\gamma_T)$
$\mathbf{c} = (\mathbf{a}_w, \mathbf{1})$	$\alpha\gamma_T$	$\beta\gamma_T + (1-\gamma_T)$
$\mathbf{c} = (\mathbf{a}_w, \mathbf{0})$	$\alpha\gamma_T + (1-\gamma_T)$	$\beta\gamma_T$

for some coefficients of optimism α and pessimism β, and a degree of ambiguity γ_T.

From Equation (16) straightforward computation yields,

$$\begin{aligned} V(a_b; D_1) &= \gamma_{10}\left[\alpha u(1) + \beta u(0)\right] + (1-\gamma_{10})\,\frac{u(1)+u(0)}{2}, \\ V(a_b; D_2) &= \gamma_{300}\left[\alpha u(1) + \beta u(0)\right] + (1-\gamma_{300})\,\frac{u(1)+u(0)}{2}. \end{aligned} \tag{17}$$

Note how the evaluation of a bet on black depends on the parameters of optimism and pessimism α and β. If the decision maker is pessimistic, i.e. $0 = \alpha < \beta \leq 1$, the evaluation of these bets will depend only on the ambiguity parameters γ_{300} and γ_{10}. Assuming that the decision maker feels more ambiguous when the decision over a bet has to be based on 10 observations than

when based on 300, i.e., $\gamma_{10} > \gamma_{300}$, one obtains

$$V(a_b; D_2) > V(a_b; D_1).$$

The reason for this result is that a pessimistic decision maker overweighs the probability of the worst outcome, 0, relative to its frequency in the data. As the number of observations increases, the weight assigned to the worst outcome diminishes. Hence, controlling for the frequency of observations, a pessimistic decision maker prefers longer data-sets.

Since the frequencies of observations in D_1 and D_2 are the same, it is easy to check that also

$$V(a_w; D_2) > V(a_w; D_1).$$

Hence, a pessimist will exhibit the usual Ellsberg preferences, choosing to bet on the urn, for which more information is available, regardless of the bet, a_w or a_b.

It is easily verified that this result obtains for any values of α and β satisfying $\beta > \alpha$, i.e. when the decision maker's degree of pessimism exceeds his degree of optimism.

If the decision maker is a pure "frequentist", i.e., if she satisfies Axiom 9A, she will evaluate data sets D independently of their length T and, hence, have a degree of ambiguity $\gamma_T = 0$ for all T. Putting $\gamma_T = 0$ in Equation (17) gives the evaluation of a bet on black a_b for an ambiguity-neutral decision maker,

$$V(a_b; D_1) = V(a_b; D_2) = \frac{1}{2}[u(1) + u(0)].$$

Hence, we will expect the frequentist to be indifferent between betting on black from Urn I (a_b, D_1) and betting on black from Urn II (a_b, D_2). Similarly, she will be indifferent between betting on white from Urn I or Urn II. ■

Example 1.1 illustrate that information in form of data sets can explain the Ellsberg two-urn paradox. Information differences regarding the urns seem to be the characteristic feature of this paradox. As ELLSBERG (1961) put it:

> "What is at issue might be called the *ambiguity* of this information, a quality depending on the amount, type, reliability, and "unanimity" of information, and giving rise to one's degree of "confidence" in an estimate of relative likelihoods". (P.657)

The amount, quality, and unanimity of information are observable characteristics of a decision

problem. It is this feature which our model can capture, thus, making at least part of the decision weights objective.

Example 2. 3 Loan market (Example 2.1 continued)

Reconsider the loan markets with entrepreneurs and the creditors facing actions from a set A and corresponding returns from R. The information about the well established Western economy is contained in the data-set D_1 and the information about the emerging Eastern European market is described by the set of observations D_2.

Specializing this model further, assume that there is only one type of project, $A = \{a\}$, say opening a fast food chain, which can result in a high financial payoff $\bar{r}$ or a low one $\underline{r}$; hence $R = \{\bar{r}; \underline{r}\}$. In order to focus on information precision, we assume that both D_1 and D_2 contain the same frequency of high payoff realizations $f := f(\bar{r}) \in (0;1)$. The data sets differ, however, in length, $T_1 := |D_1| > |D_2| = T_2$.

For a contracted repayment amount $q \in (\underline{r}; \bar{r})$, the payoff of a lender is given by the minimum of q and $\underline{r}$, $\min\{q, \underline{r}\}$, and the payoff of an entrepreneur is $\max\{\bar{r} - q, 0\}$. Given the information in a data set $D \in \{D_1, D_2\}$, one obtains from the representation in Equation (11),

$$\begin{aligned} V_L(a;D) &= u(q)\, h^L(\bar{r}|a,D) + u(\underline{r})\, h^L(\underline{r}|a,D) \\ &= u(q)\left[f\hat{p}_{T,a}^{(a,\bar{r})}(\bar{r}) + (1-f)\,\hat{p}_{T,a}^{(a,\underline{r})}(\bar{r})\right] + u(\underline{r})\left[f\hat{p}_{T,a}^{(a,\bar{r})}(\underline{r}) + (1-f)\,\hat{p}_{T,a}^{(a,\underline{r})}(\underline{r})\right] \\ &= \gamma_T^L\left[\alpha^L u(q) + \beta^L u(\underline{r})\right] + \left(1-\gamma_T^L\right)\left[fu(q) + (1-f)\,u(\underline{r})\right] \\ &= \gamma_T^L \cdot \left[\left(1-\beta^L\right)u(q) + \beta^L u(\underline{r})\right] + \left(1-\gamma_T^L\right)\left[fu(q) + (1-f)\,u(\underline{r})\right] \end{aligned}$$

for the lender, and, similarly, for the entrepreneur,

$$\begin{aligned} V_E(a;D) &= u(\bar{r}-q)\, h^E(\bar{r}|a,D) + u(0)\, h^E(\underline{r}|a,D) \\ &= \gamma_T^E \cdot \left[\left(1-\beta^E\right)u(\bar{r}-q) + \beta^E u(0)\right] + \left(1-\gamma_T^E\right)\left[fu(\bar{r}-q) + (1-f)\,u(0)\right]. \end{aligned}$$

In order to focus on the degree of ambiguity γ_T^i and the optimism and pessimism parameters α^i, β^i, we assumed identical utilities over the outcomes $u^i(\cdot) = u(\cdot)$ for the agents $i = E, L$. In addition, we used the fact that $\alpha^i + \beta^i = 1$ holds in order to eliminate the parameter $\alpha^i = 1 - \beta^i$. Consequently, a pure optimist corresponds to $\beta^i = 0$ and a pure pessimist to $\beta^i = 1$.

Depending on the relative magnitude of the parameters (β^i, γ_T^i), preferences about in which market to undertake the action a may differ. Notice however, that if the agents share the degree

of pessimism, $\beta^i = \beta$, and assess the ambiguity by the same degree $\gamma_T^i = \gamma_T$; for $i = E, L$, both the entrepreneur and the lender will unanimously prefer investing in the market with the better information, $T_1 > T_2$, and hence the lower degree of ambiguity, $V_i(a; D_1) > V_i(a; D_{21})$ provided the degree of pessimism is sufficiently high, i.e., β is close to 1. If $\beta = 1$, this would be true without reference to the payoffs, since the expected utility over outcomes with respect to the frequency f is always larger than the utility of the smallest payoff.

More interesting is the case, when entrepreneurs and lenders agree with respect to the ambiguity involved with investing in the respective markets, $\gamma_{T_1}^i = \gamma_{T_1} < \gamma_{T_2} = \gamma_{T_{21}}^i$, but differ in their attitude towards ambiguity. Assuming that lenders will be more conservative than entrepreneurs, for convenience of computation, we will consider the extreme case where the entrepreneurs are pure optimists and the lenders pure pessimists, $0 = \beta^E < \beta^L = 1$. Lenders will provide funding for entrepreneurs investing in both markets only if the contracted return q_2 in the market with the more ambiguous information D_2 is sufficiently higher than the one contracted for the less ambiguous one, D_1. Assuming that there are no other investment opportunities an equilibrium system of contracted returns (q_1^*, q_2^*) must be such that

$$V_L(a; D_1|q_1^*) = V_L(a; D_2|q_2^*) \quad \textit{and} \quad V_E(a; D_1|q_1^*) = V_E(a; D_2|q_2^*)$$

holds. Hence, one obtains

$$\begin{aligned} & \gamma_{T_1}^L u(\underline{r}) + \left(1 - \gamma_{T_1}^L\right)\left[f u(q_1) + (1-f)\, u(\underline{r})\right] \\ = \; & \gamma_{T_2}^L u(\underline{r}) + \left(1 - \gamma_{T_2}^L\right)\left[f u(q_2) + (1-f)\, u(\underline{r})\right], \\ & \gamma_{T_1}^E u(\overline{r} - q_1) + \left(1 - \gamma_{T_1}^E\right)\left[f u(\overline{r} - q_1) + (1-f)\, u(0)\right] \\ = \; & \gamma_{T_2}^E u(\overline{r} - q_2) + \left(1 - \gamma_{T_2}^E\right)\left[f u(\overline{r} - q_{21}) + (1-f)\, u(0)\right], \end{aligned}$$

For the case of a linear utility function, $u(x) = x$, straightforward computations yield the explicit solution

$$q_1^* = \frac{\left(\gamma_{T_1} - \gamma_{T_2}\right)\left[\left(1 - \gamma_{T2}\right)(1-f)\,\overline{r} + \left(f + \gamma_{T2}(1-f)\right) f\underline{r}\right]}{\left(1 - \gamma_{T_2}\right)\left(\gamma_{T_1} + (1-\gamma_{T_1})f\right) - \left(1 - \gamma_{T_1}\right)\left(\gamma_{T_2} + (1-\gamma_{T_2})f\right) f}$$

and

$$q_2^* = \frac{q_1^*\left(1 - \gamma_{T_1}\right) + \underline{r}\left(\gamma_{T_1} - \gamma_{T_2}\right)}{1 - \gamma_{T_2}} f.$$

It is easy to check that $q_2^* \in (\underline{r}; \overline{r})$, while $q_1^* \in (\underline{r}; q_2^*)$.

Hence, the cost of credit is lower in the more informative market. If one would compare the empirical distribution of returns without taking into account the informativeness of data, one would expect that both markets would be served at the same price, i.e., $\gamma_{T_1} = \gamma_{T_2}$ *implies* $q_1^* = q_2^*$. ■

Example 2.1 illustrates the potential of this approach to generate new and testable hypothesis about market outcomes which could not be obtained with expected utility theory. Given the same frequencies of outcomes, the Bayesian model would predict that a decision maker with a given prior for the probability of high returns would prefer to trade in the market with more observations, D_1, if the realized frequency f of the better outcome would exceed her prior. Otherwise, she would want to trade in the market D_2. In contrast, a purely pessimistic decision maker will always choose to trade in the more informative market, regardless of the realized frequency f, provided that the price of credit in both markets is equal.

Our next example illustrates the impact of similarity on the decision maker's evaluation of different information contexts.

Example 3. 3 Financial Investment (Example 3.1 continued)

Suppose that the investor is faced with the choice of investing in a listed internet retail company in his home market or in a foreign market. Thus, he wants to compare $(a_1; D_H)$ *to* $(a_1; D_F)$*, where* D_H *and* D_F *are given by (*3*) and (*4*). To isolate the effect of quality of information of the investor's decision, we assume that the performance of* a_1 *in both markets is identical, i.e.* $r_t^1 = \tilde{r}_t^1$ *for all* $t \in \{1...T\}$*. Let*

$$f_i(r) =: \frac{\{(a_i; r_t^i) \mid t \in \{1...T\}, r_t^i = r\}}{T}$$

be the frequency of returns generated by asset $i \in \{1, 2\}$*.*

According to our representation, the investor's preferences will be determined by the compari-

son of

$$\begin{aligned} & V(a_1; D_H) \\ = & \sum_{\tilde{r} \in R} u(\tilde{r}) \left[\frac{\sum_{c \in C} s_{a_c}(a_1) f_{D_H}(c) \left[(1-\gamma_{2T}) h^c_{a_1}(\tilde{r}) + \gamma_{2T} (\alpha \delta_{\bar{r}}(\tilde{r}) + \beta \delta_{\underline{r}}(\tilde{r})) \right]}{\sum_{c \in C} s_{a_1}(a_2) f_{D_H}(c)} \right] \\ = & \sum_{\tilde{r} \in R} u(\tilde{r}) \left[\frac{\sum_{r \in R} (1-\gamma_{2T}) \left[f_1(r) \delta_r(\tilde{r}) + s_{a_1}(a_2) f_2(r) h^{(a_2;r)}_{a_1}(\tilde{r}) \right]}{1 + s_{a_1}(a_2)} + \gamma_{2T} [\alpha \delta_{\bar{r}}(\tilde{r}) + \beta \delta_{\underline{r}}(\tilde{r})] \right] \end{aligned}$$

and

$$\begin{aligned} & V(a_1; D_F) \\ = & \sum_{\tilde{r} \in R} u(\tilde{r}) \left[\frac{\sum_{c \in C} s_{a_c}(a_1) f_{D_F}(c) \left[(1-\gamma_T) h^c_{a_1}(\tilde{r}) + \gamma_T (\alpha \delta_{\bar{r}}(\tilde{r}) + \beta \delta_{\underline{r}}(\tilde{r})) \right]}{\sum_{c \in C} s_{a_1}(a_2) f_{D_F}(c)} \right] \\ = & \sum_{\tilde{r} \in R} u(\tilde{r}) \left[\sum_{r \in R} (1-\gamma_T) f_1(r) \delta_r(\tilde{r}) + \gamma_T [\alpha \delta_{\bar{r}}(\tilde{r}) + \beta \delta_{\underline{r}}(\tilde{r})] \right]. \end{aligned}$$

Here, γ_T and γ_{2T} with $\gamma_T > \gamma_{2T}$ reflect the perceived degree of ambiguity for the two data-sets, D_F, which contains T observations and D_H containing $2T$ observations. α and $\beta = 1 - \alpha$ are the investor's degrees of optimism and pessimism. $s_{a_1}(a_2) \in (0;1)$ is the perceived similarity between the listed company a_1 and the non-listed one, a_2. For the purposes of the example, we assume that the decision maker is purely pessimistic, i.e. $\alpha = 0$, $\beta = 1$.

The interpretation is as follows: the fact that companies a_1 and a_2 operate in the same industry implies that the investor might consider the information about the returns of a_2 to be relevant for his prediction about the returns of a_1. For instance, he might believe that the returns in a given industry within a country are very highly correlated, implying $h^{(a_2;r)}_{a_1} = \delta_r$. However, he might assign the evidence obtained from a_2 a weight smaller than 1 when using it to predict the performance of a_1. This is reflected in the similarity weight $s_{a_1}(a_2)$ assigned to the cases containing a_2. Since D_H contains a larger number of observations than D_F, the investor perceives a smaller degree of ambiguity with respect to D_H, and being purely pessimistic, assigns a smaller weight on the worst possible outcome $\underline{r}$ in the home market (γ_{2T}) than in the foreign market (γ_T).

Since the performance of a_1 is identical for the two markets, investor's preferences will be determined by the empirical frequency of returns of a_2 and by the perceived similarity between a_1

and a_2. The investor will prefer to invest in the home market if:

$$V(a_1; D_H) \geq V(a_1; D_F),$$

or if

$$\frac{s_{a_1}(a_2)}{1+s_{a_1}(a_2)} \sum_{\tilde{r}\in R} u(\tilde{r}) \sum_{r\in R} \left[f_2(r)\, h_{a_1}^{(a_2;r)}(\tilde{r}) - f_1(r)\, \delta_r(\tilde{r}) \right] \tag{18}$$
$$+\frac{\gamma_T - \gamma_{2T}}{1-\gamma_{2T}} \sum_{\tilde{r}\in R} u(\tilde{r}) \sum_{r\in R} \left[f_1(r)\, \delta_r(\tilde{r}) - \delta_{\underline{r}}(\tilde{r}) \right] \geq 0$$

*The first term in (*18*) represents the effect that the additional information about a_2 has on the unambiguous forecast about a_1 in the home market. The size of this effect will depend on the similarity between a_1 and a_2. Its sign will depend on whether the additional evidence obtained from a_2 is considered to be more favorable for the choice of a_1 than the evidence based on the set of observations containing a_1, i.e., whether*

$$\sum_{\tilde{r}\in R} u(\tilde{r}) \sum_{r\in R} \left[f_2(r)\, h_{a_1}^{(a_2;r)}(\tilde{r}) - f_1(r)\, \delta_r(\tilde{r}) \right] \geq 0 \tag{19}$$

*The second term in (*18*) reflects the fact that the additional information about a_2 decreases the perceived ambiguity and reduces the weight the decision maker assigns to the worst possible outcome in his evaluation of a_1. Since $\gamma_T > \gamma_{2T}$, for a purely pessimistic decision maker the second effect is always positive. When (*19*) holds, investing in the home market is strictly preferred. However, the investor might prefer the home market, even if (*19*) is violated, provided that the relative decrease in ambiguity $\frac{\gamma_T - \gamma_{2T}}{1-\gamma_{2T}}$ is sufficiently large and the relevance of a_2 for the evaluation of a_1, $s_{a_1}(a_2)$ is sufficiently small.*

We conclude that the investor's preferences for home assets need not be based on superior performance of home assets or on the availability of highly relevant information about other home assets. Rather, the fact that a pessimistic investor has access to a larger and somewhat relevant set of observations from his home market renders home assets more preferred and leads him to exhibit behavior consistent with the well-known home bias. ■

The following example which resumes Example 1.1 provides an opportunity to illustrate the potential of our approach for dealing with the value of information. It is beyond the scope of this paper to deal with the problem of valuing information in general. A weak notion of value of information consists in its capacity to improve decisions. In particular, information has some

value for the decision maker if it changes the evaluation of at least one act.

Example 4. 3 Medical treatment (Example 4.1 continued)

Reconsider the medical doctor who needs to choose a treatment from a set of actions A given her information D. As before, outcomes are elements of the set R and data-bases are elements of $\mathbb{D}$. Suppose there is a new study which offers a data set D' for the new drug, i.e., the treatment option a_1. The new study provides valuable information if it leads to a new assessment of the new drug, i.e., if $V(a_1, D \cup D') \neq V(a_1, D)$.

Recall that $h^c_{a_1}$ denotes the limit distribution over outcomes of action a_1 based on observations of cases (a_c, r_c) with $a_c \in A$ and $r_c \in R$. In order to simplify notation denote by

$$h^D_{a_1} =: \sum_{c \in C} \frac{s_{a_1}(a_c) f_D(c)}{\sum_{c \in C} s_{a_1}(a_c) f_D(c)} h^c_{a_1} \tag{20}$$

the unambiguous prediction for the distribution over outcomes of action a_1 based on data-set D. This allows us to write

$$\begin{aligned} V(a_1, D) &= \sum_{r \in R} u(r) h(r|a_1, D) \\ &= \sum_{r \in R} u(r) \left[(1 - \gamma_T) h^D_{a_1}(r) + \gamma_T (\alpha \delta_{\bar{r}}(r) + \beta \delta_{\underline{r}}(r)) \right] \end{aligned}$$

and

$$\begin{aligned} V(a_1, D \cup D') &= \sum_{r \in R} u(r) h(r|a_1, D \cup D') \\ &= \sum_{r \in R} u(r) \left[\left(1 - \gamma_{T+T'}\right) h^{D \cup D'}_{a_1}(r) + \gamma_{T+T'} (\alpha \delta_{\bar{r}}(r) + \beta \delta_{\underline{r}}(r)) \right] \end{aligned}$$

for the combined data-sets D and D'.

It is straightforward to check that

$$\begin{aligned} & \left(1 - \gamma_{T+T'}\right) h^{D \cup D'}_{a_1} + \gamma_{T+T'} (\alpha \delta_{\bar{r}} + \beta \delta_{\underline{r}}) \\ = \; & \left(1 - \gamma_{T+T'}\right) \left[s h^D_{a_1} + (1 - s) h^{D'}_{a_1} \right] + \gamma_{T+T'} (\alpha \delta_{\bar{r}} + \beta \delta_{\underline{r}}), \end{aligned}$$

with the positive weight

$$s := \frac{T \sum_{c \in C} s_{a_1}(a_c) f_D(c)}{T \sum_{c \in C} s_{a_1}(a_c) f_D(c) + T' \sum_{c \in C} s_{a_1}(a_c) f_{D'}(c)}.$$

Hence, one has

$$\begin{aligned} & V(a_1, D \cup D') - V(a_1, D) \\ = & \sum_{r \in R} u(r) \Big\{ \left(1 - \gamma_{T+T'}\right) \left[s h^D_{a_1}(r) + (1-s)\, h^{D'}_{a_1}(r) \right] + \gamma_{T+T'} \left(\alpha \delta_{\bar{r}}(r) + \beta \delta_{\underline{r}}(r) \right) \\ & - \left(1 - \gamma_T\right) h^D_{a_1}(r) + \gamma_T \left(\alpha \delta_{\bar{r}}(r) + \beta \delta_{\underline{r}}(r) \right) \Big\} \\ = & \sum_{r \in R} u(r) \left\{ s \left(1 - \gamma_{T+T'}\right) \left[h^D_{a_1}(r) - h^{D'}_{a_1}(r) \right] + \left(\gamma_{T+T'} - \gamma_T \right) \left[\alpha \delta_{\bar{r}}(r) + \beta \delta_{\underline{r}}(r) - h^{D'}_{a_1}(r) \right] \right\}. \end{aligned}$$

For arbitrary data-sets D and D', the additional information of data set D' will be without value for the doctor, $V(a_1, D \cup D') - V(a_1, D) = 0$, if

• *there is no difference in the unambiguous limiting distributions over outcomes resulting from the two data-sets, $h^D_{a_1}(r) = h^{D'}_{a_1}(r)$, and if*

• *the additional data would not increase the doctor's confidence in the unambiguous prediction, $\gamma_{T+T'} = \gamma_T$.*

■

Example 4.21 illustrates minimal conditions for information from a new additional data-set to be of value for the evaluation of the treatment a_1. Clearly, for information to be valuable in decision making it must affect the evaluation of at least one action. One could think of other, more restrictive notions of "value of information", e.g., that the actual choice of action should be changed by new information.

Notice that these notions of value for a data set do not depend on how the actions are evaluated. For a medical study, one certainly would not want to know just favorable evidence for a new drug. Unfavorable evidence would be equally valuable, since it may lead to a better treatment. On the other hand, our model of evaluation of data bases would allow us to consider preferences for "particular types of information". This need not be just "wishful thinking", since information, though useful for treatment, may make a patient "feel worse". This is a further field which can be modelled and explored with the approach suggested in this paper.

6 Related Literature

Our paper connects the model of case-based decision making to the literature on ambiguity-

aversion. While several axiomatizations of case-based decision making have been provided in the literature, see for instance Gilboa and Schmeidler (1997, 2001) and Gilboa, Schmeidler and Wakker (2002), the representation suggested by BGSS (2005), which in our view is best suited to model learning from data, has not been axiomatized up to now. We close this gap by providing such an axiomatization.

Furthermore, we extend the BGSS (2005) approach by allowing the decision maker to take into account ambiguity. There are several ways to model ambiguity in the literature: a representation of ambiguous beliefs by means of capacities was introduced by Schmeidler (1989). Bewley (1986)'s approach is based on incomplete preferences. The paper of Klibanoff, Marinacci and Mukerji (2005) models ambiguity as second-order risk with respect to the probability distribution determining the outcome. The multiple-prior approach developed by Gilboa and Schmeidler (1989) generalized by Ghirardato, Maccheroni and Marinacci (2004) and Chateauneuf, Eichberger and Grant (2007) represents ambiguity by a set of probability distributions which a decision maker considers when evaluating her expected utility. In all of these models, ambiguity is subjective, i.e. it is derived only from the preferences of the decision maker and is not based on the presence (or absence) of objective information.

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, which allows them to separate the objectively given Knightian uncertainty from the subjective attitude towards ambiguity.

In our approach, it appears natural to use the length of the data-set (an objective characteristic) to measure the ambiguity of the available information. In contrast, the subjective perception of ambiguity and the attitude towards ambiguity may differ across subjects. We can identify these by asking decision makers to compare not only actions, but also the evidence that different data-sets give in support of a given action. By imposing a relatively mild condition on such preferences, we are able to separate the perception of ambiguity from the attitude towards ambiguity. The representation of beliefs we obtain is very similar to the one used by Chateauneuf,

Eichberger and Grant (2007). In this sense, our paper is also closely related to the approach of Gonzales and Jaffray (1998). They model preferences over Savage-type acts for a given set of, possibly imprecise, data. They derive a representation of preferences in form of a linear combination of the maximal and minimal potential outcome of an act and its expected utility with respect to the observed frequency of states. The weights attached to the maximal and minimal outcomes can be interpreted as degrees of optimism and pessimism. They decrease over time relative to the weight attached to the expected utility part of the representation. Because observations may be imprecise a decision maker associates with a set of data a set of priors centered around the observed frequency. The size of the set of probabilities depends negatively on the amount of data. We are able to derive a similar representation, but without making use of the rich structure of the Savage framework. Furthermore, our approach encompasses a richer class of situations which allow for, but are not restricted to, the case of controlled statistical experiments considered in both Coignard and Jaffray (1994) and Gonzales and Jaffray (1998). The data-sets in our framework consist of heterogenous cases and allow us to integrate perceptions of similarity into the decision making process.

Our paper provides a model of learning which is very different from the Bayesian, as axiomatized by Easley and Rustichini (1999). First, even in the case of ambiguity-neutrality, the behavior we model is closer to the frequentist rather than the Bayesian approach: in particular, observing a data-set in which action a pays the highest possible outcome with frequency 1 would lead the decision maker in our model to assign a probability 1 to this outcome, regardless of the number of observations, making his behavior inconsistent with the one of a Bayesian. This is due to the fact that the decision maker is ambiguity-neutral, hence a change in the degree of model uncertainty he experiences does not influence his choices. In the case of ambiguity-aversion, the probability distribution over outcomes is no longer independent of the number of observations. However, a comparison of the probability distributions over outcomes assigned to actions which are bets on mutually exclusive events, will in general imply that the decision maker does not entertain a probability distribution over these events. Hence, even though an

ambiguity-averse decision maker reacts to reduction in model uncertainty, he does so in a very different way from a Bayesian decision maker.

Epstein and Schneider (2007) analyze statistical learning in the context of ambiguity. Their approach distinguishes between two types of scenarios: those, in which it is possible to learn the objective probability distribution over outcomes and, thus, in the limit the ambiguity disappears, and scenarios, in which the ambiguity persists even in the limit. Our framework distinguishes between controlled statistical experiments, and situations in which relevant, but not completely identical cases have been observed. We show that the decision maker will be able to learn the objective probability distribution in controlled statistical experiments when the size of the data-set becomes large. In this case, ambiguity disappears and the subjective beliefs of the decision maker exactly coincide with the observed empirical frequencies. When cases are distinct from the situation under consideration, the limit prediction is subjective and might be very different from the observed frequency of outcomes. This is due to the fact that the decision maker does not observe counterfactuals, and, hence, cannot learn anything about the correlation across different actions. Hence, the subjectivity of the limit probability distribution can be attributed to ambiguity related to this lack of information.

7 Conclusion

In this paper, we analyzed preferences among actions characterized by an information context. We derived an expected utility representation for this situation, separating utility over outcomes from subjective beliefs. We obtained a separation of the beliefs into three components: the frequency of cases in the data-set, the relevance of each of the cases for the prediction to be made and the probabilistic prediction associated with each case. For the case of controlled statistical experiments, we showed that an ambiguity-neutral decision maker will use the frequency of observed outcomes to make predictions. However, the assumption of ambiguity-neutral behavior represents only a special case of our formalism. Allowing preferences of the decision maker to depend on the length of the data-set, we are able to identify the subjectively perceived degree

of ambiguity and to divide it into two components: the degree of optimism and the degree of pessimism. While the perceived ambiguity declines as the number of observations increases, the relative degrees of optimism and pessimism are constant across data-sets. Our examples illustrate the scope of phenomena which can be described using our approach. We show how the Ellsberg paradox can be generalized to apply to different degrees of information precision. More generally, we demonstrate how preferences across different information contexts are influenced by the number and by the type of observations. Our results can be applied to model agents' choice of a market in which they would like to participate, and to examine phenomena related to the well-documented home bias. Finally, our approach provides a framework, which allows us to speak of the value of additional information.

8 Appendix

Proof of Theorem 4.1:

The non-triviality condition implies that there exist a, c and c' such that:

$$\left(a; c^T\right) \succ \left(a; \left(c'\right)^T\right)$$

for some $T \in \mathbb{N}$. Then, by Axiom 6,

$$\left(a; \left(a; \bar{r}\right)^T\right) \succsim \left(a; c^T\right) \succ \left(a; \left(c'\right)^T\right) \succsim \left(a; \left(a; \underline{r}\right)^T\right).$$

Hence,

$$\left(a; \left(a; \bar{r}\right)^T\right) \succ \left(a; \left(a; \underline{r}\right)^T\right).$$

By Axiom 3, this holds for all $T \in \mathbb{N}$.

Note that every function $V : A \times \mathbb{D} \to \mathbb{R}$ which represents $\succsim$ in the sense of (11) is bounded. Alternatively, take a real-valued function U defined on $A \times \mathbb{D}$ representing $\succsim$ (but not necessarily in the sense of (11)). Since the set $A \times \mathbb{D}$ is countable, such a representation always exists. Furthermore, there exists an order-preserving mapping of the values of U to $\mathbb{Z} \setminus \{0\}$. The function $U : A \times \mathbb{D} \to \mathbb{Z} \setminus \{0\}$ can then be transformed by a monotone transformation into

$$V = \begin{cases} 1 - \frac{1}{2(1+U)} & \text{if } U > 0 \\ \frac{1}{2U} & \text{if } U < 0 \end{cases}$$

to obtain a function that is bounded between $[0;1]$. Note that V also represents $\succsim$. Hence, whenever a representation of $\succsim$ exists, there is also a bounded function V which represents $\succsim$. We, therefore, construct a bounded function V, which satisfies the properties specified above. Note that if V represents $\succsim$, we must have by Axiom 10,

$$V\left(a;(a;\bar{r})^{T+1}\right) \geq V\left(a;(a;\bar{r})^{T}\right)$$
$$V\left(a;(a;\underline{r})^{T}\right) \geq V\left(a;(a;\underline{r})^{T+1}\right)$$

for all T. Also, by Axiom 10, the sequences $V\left(a;(a;\bar{r})^{T}\right)$ and $V\left(a;(a;\underline{r})^{T}\right)$ are either strictly monotone or constant. Since V is bounded, they converge. Since

$$\left(a;(a;\bar{r})^{T}\right) \succ \left(a;(a;\underline{r})^{T}\right)$$

holds for all T, it follows that the two limits are distinct. By Axiom 6, the limits satisfy:

$$\lim_{T\to\infty} V\left(a;(a;\bar{r})^{T}\right) \geq V(a;D) \geq \lim_{T\to\infty} V\left(a;(a;\underline{r})^{T}\right).$$

We, therefore, set:

$$\lim_{T\to\infty} V\left(a;(a;\bar{r})^{T}\right) = :1$$
$$\lim_{T\to\infty} V\left(a;(a;\underline{r})^{T}\right) = :0$$

Take a data-set $(a;r)$. By Axiom 7, we can find a sequence $(\mu_T)_{T\in\{1!,2!\ldots n!\ldots\}}$ with $\mu_T \in [0;1]$ such that

$$\mu_T = \min_{m\in\left\{\frac{1}{T};\frac{2}{T}\ldots 1\right\}} \left(a;\left(m\delta_{(a;\bar{r})} + (1-m)\,\delta_{(a;\underline{r})};T\right)\right) \succsim (a;(a;r))$$

and similarly, a sequence $(\nu_T)_{T\in\{1!,2!\ldots n!\ldots\}}$, $\nu_T \in [0;1]$ such that:

$$\nu_T = (a;(a;r)) \succsim \max_{n\in\left\{\frac{1}{T};\frac{2}{T}\ldots 1\right\}} \left(a;\left(n\delta_{(a;\bar{r})} + (1-n)\,\delta_{(a;\underline{r})};T\right)\right).$$

Since $\left(a;(a;\bar{r})^{T}\right) \succ \left(a;(a;\underline{r})^{T}\right)$, Axiom 4 implies that:

$$\left(a;\left(m\delta_{(a;\bar{r})} + (1-m)\,\delta_{(a;\underline{r})};T\right)\right) \succ \left(a;\left(n\delta_{(a;\bar{r})} + (1-n)\,\delta_{(a;\underline{r})};T\right)\right),$$

whenever $m > n$. Hence, unless $\mu_T = \nu_T$, it must be that $\mu_T - \nu_T = \frac{1}{T}$. Since $\mu_T - \nu_T \leq \frac{1}{T}$ for all T, we have that $\lim_{T\to\infty}(\mu_T - \nu_T) = 0$. Hence, in the limit, we can construct a data-set which contains only observations of the type $(a;\bar{r})$ and $(a;\underline{r})$ and is considered to be exactly

indifferent to $(a;r)$. By Axiom 9, there is an outcome $\hat{r}$ such that

$$(a;(a;\hat{r})) \sim \left(a;(a;\hat{r})^k\right)$$

for any $k \in \mathbb{N}$. Suppose first that $(a;(a;r)) \sim (a;(a;\hat{r}))$, then μ_T and ν_T will be such that:

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right) \succsim (a;(a;r)) \sim (a;(a;\hat{r})) \succsim \left(a;\left(\nu_T\delta_{(a;\bar{r})} + (1-\nu_T)\,\delta_{(a;\underline{r})};T\right)\right).$$

Axiom 10 implies that as long as

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right) \succ (a;(a;\hat{r})),$$

or

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T+1\right)\right) \succ \left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right).$$

the sequence μ_T is decreasing. If for some $\bar{T}$,

$$\left(a;(a;\bar{r})^{\mu_T T} \circ a;(a;\underline{r})^{(1-\mu_T)T}\right) \sim (a;(a;\hat{r})),$$

then $\mu_T = \mu_{\bar{T}}$ for all $T \geq \bar{T}$. By a similar argument, the sequence ν_T is either constant or increasing. Since furthermore $\lim_{T\to\infty}(\mu_T - \nu_T) = 0$, it follows that both sequences converge to the same limit. Denote this limit by $\hat{\mu}$.

Now suppose that $(a;(a;r)) \succ (a;(a;\hat{r}))$, then for sufficiently large T's, μ_T and ν_T will be such that:

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right) \succsim (a;(a;r)) \succsim \left(a;\left(\nu_T\delta_{(a;\bar{r})} + (1-\nu_T)\,\delta_{(a;\underline{r})};T\right)\right) \succ (a;(a;\hat{r})).$$

Since

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right) \succ (a;(a;\hat{r})),$$

by Axiom 10, we have:

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T+1\right)\right) \succ \left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right)$$

if

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right) \succ (a;(a;r))$$

and

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T+1\right)\right) \sim \left(a;\left(\mu_T\delta_{(a;\bar{r})} + (1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right)$$

if

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})}+(1-\mu_T)\,\delta_{(a;\underline{r})};T+1\right)\right)\sim(a;(a;r))\,.$$

It follows that the sequence μ_T is either constant or increasing. Since it is bounded, it has a limit and since $\lim_{T\to\infty}(\mu_T-\nu_T)=0$, it follows that μ_T and ν_T converge to the same limit, denoted by $\mu^a_{(a;r)}$. A symmetric argument applies for the case, in which $(a;(a;\hat{r}))\succ(a;(a;r))$. Obviously, the same construction can be repeated for any data-set $D\in\mathbb{D}$. Denote the common limit of the corresponding sequences μ_T and ν_T by μ^a_D.

Now define the function $V(a;D):\mathbb{D}\to[0;1]$ by:

$$V(a;D)=:\mu^a_D.$$

To see that the function $V(a;D)$ represents $\succsim$, suppose that $(a;D)\succsim(a;D')$. Then, by construction,

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})}+(1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right)\succsim\left(a;\left(\mu'_T\delta_{(a;\bar{r})}+(1-\mu'_T)\,\delta_{(a;\underline{r})};T\right)\right)$$

for all sufficiently large T. Axioms 3, 4 and 6 imply that

$$\left(a;\left(\mu_T\delta_{(a;\bar{r})}+(1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right)\succsim\left(a;\left(\mu'_T\delta_{(a;\bar{r})}+(1-\mu'_T)\,\delta_{(a;\underline{r})};T\right)\right)$$

if and only if $\mu_T\geq\mu'_T$. Since this holds for all sufficiently large T's, it follows that the limits of the sequence also satisfy $\mu^a_D\geq\mu^a_{D'}$. By construction, this is equivalent to

$$V(a;D)\geq V(a;D')\,.$$

Furthermore, Axiom 5 implies that for $D\in\mathbb{D}_a$, $D'\in\mathbb{D}_{a'}$ such that $f_D(a;r)=f_{D'}(a';r)$ for all $r\in R$, $\mu^a_D=\mu^{a'}_{D'}$, and hence, $V(a;D)=V(a';D)$. It follows that for any $D\in\mathbb{D}_a$, $D'\in\mathbb{D}_{a'}$,

$$(a;D)\succsim(a';D')$$

if and only if

$$(a;D)\succsim\left(a;\tilde{D}'\right),$$

where $\tilde{D}'\in\mathbb{D}_a$ is such that $f_{D'}(a';r)=f_{\tilde{D}'}(a;r)$ for all $r\in R$, i.e. if and only if

$$\mu^a_D\geq\mu^a_{\tilde{D}'}=\mu^a_{D'}.$$

Finally, take any two data-sets D and D'. Then $(a;D) \succsim (a';D')$ implies that:

$$\begin{aligned}\left(a;\left(\mu_T\delta_{(a;\bar{r})}+(1-\mu_T)\,\delta_{(a;\underline{r})};T\right)\right) &\succsim \left(a';\left(\mu'_T\delta_{(a';\bar{r})}+(1-\mu'_T)\,\delta_{(a';\underline{r})};T\right)\right)\\ &\sim \left(a;\left(\mu'_T\delta_{(a;\bar{r})}+(1-\mu'_T)\,\delta_{(a;\underline{r})};T\right)\right)\end{aligned}$$

for all sufficiently large values of T. This is equivalent to

$$\mu_D^a \geq \mu_{D'}^{a'},$$

or, by construction, to $V(a;D) \geq V(a';D')$.

For each r, let λ_r be such that:

$$V\left(a;(a;r)^T\right) = \lambda_r V\left(a;(a;\bar{r})^T\right) + (1-\lambda_r)\,V\left(a;(a;\underline{r})^T\right).$$

Such a $\lambda_r \in [0;1]$ exists by Axiom 6. By Axioms 5 and 8, this λ_r is independent of a and T. Hence, we can define the von-Neumann-Morgenstern utility index $u(r)$ to be

$$u(r) =: \lambda_r$$

for all $a \in A$ and $r \in R$. Obviously, $u(\bar{r}) = 1$ and $u(\underline{r}) = 0$. Note that by construction, for all $r \in R$,

$$\lim_{T\to\infty} V\left(a;(a;r)^T\right) = u(r).$$

For each $V(a;D) = \mu_D^a$ and for a given von-Neumann-Morgenstern utility function u, there exist a set of probability distributions over outcomes, $h \in \Delta^{|R|-1}$ such that

$$h \cdot u = \mu_D^a.$$

Denote this set by

$$H_a(D) =: \left\{h \in \Delta^{|R|-1} \mid h \cdot u = \mu_D^a\right\}$$

and note that, e.g., the probability distribution h with $h(\bar{r}) = \mu_D^a$ and $h(\underline{r}) = (1-\mu_D^a)$ is an element of this set.

Consider a data-set $D \in \mathbb{D}_{a'}^T$. Using the construction above, for all $r \in R$ such that $f_D(a';r) > 0$, we can find sequences $\left(\mu_{\hat{T}}^r\right)$ and $\left(\nu_{\hat{T}}^r\right)$ such that:

$$\left(a;\left(\mu_{\hat{T}}^r\delta_{(a;\bar{r})}+\left(1-\mu_{\hat{T}}^r\right)\delta_{(a;\underline{r})};\hat{T}\right)\right) \succsim \left(a;\left(a';r\right)^T\right) \succsim \left(a;\left(\nu_{\hat{T}}^r\delta_{(a;\bar{r})}+\left(1-\nu_{\hat{T}}^r\right)\delta_{(a;\underline{r})};\hat{T}\right)\right).$$

Applying iteratively Axiom 4, we conclude that:

$$\left(a;\left(\sum_{r\in R}\left(\mu^r_{\hat{T}}\delta_{(a;\bar{r})}+\left(1-\mu^r_{\hat{T}}\right)\delta_{(a;\underline{r})}\right)f_D\left(a';r\right);\hat{T}\right)\right)\succsim\left(a;D\right)$$
$$\succsim\left(a;\left(\sum_{r\in R}\left(\nu^r_{\hat{T}}\delta_{(a;\bar{r})}+1-\nu^r_{\hat{T}}\delta_{(a;\underline{r})}\right)f_D\left(a';r\right);\hat{T}\right)\right).$$

It follows that the limits of the corresponding sequences satisfy:

$$\mu^a_D=\sum_{r\in R}\mu^a_{(a';r)^T}f_D\left(a';r\right).$$

Hence, for any $D\in\mathbb{D}^T_{a'}$, we have:

$$V\left(a;D\right)=\sum_{r\in R}V\left(a';r\right)^T f_D\left(a';r\right).$$

It follows immediately that for any $D\in\mathbb{D}^T_{a'}$,

$$V\left(a;D\right)=\sum_{r\in R}u\left(r\right)h\left(\left(a';r\right)^T\right)f_D\left(a';r\right)$$

for some $h\left(a';r\right)^T\in H_a\left(\left(a';r\right)^T\right)$.

Take three data-sets D, D' and $D''\in\mathbb{D}^{T'}$ with corresponding frequencies f, f' and f'' such that $\eta f+\left(1-\eta\right)f''=f'$ for some $\eta\in\left(0;1\right)$. By Axiom 3, $\left(f;T\right)\succ\left(f'';T\right)$ implies

$$\left(f;kT\right)\succ\left(f';kT\right)\succ\left(f'';kT\right)$$

for all $k\in\mathbb{N}$. Hence, $\mu^a_D>\mu^a_{D'}>\mu^a_{D''}$ and, furthermore, there is a $\lambda\in\left(0;1\right)$ such that

$$\mu^a_{D'}=\lambda\mu^a_D+\left(1-\lambda\right)\mu^a_{D''},$$

and, by Axiom 8,

$$\mu^a_{(D')^k}=\lambda\mu^a_{D^k}+\left(1-\lambda\right)\mu^a_{(D'')^k}$$

for all $k\in\mathbb{N}$. Applying the claim inductively, we obtain that for any frequencies $\left(f_i\right)_{i=1}^n$ and f in F^T such that $\sum_{i=1}^n\eta_i f_i=f$, for some coefficients $\eta_i\in\left(0;1\right)$ with $\sum_{i=1}^n\eta_i=1$, there are coefficients $\left(\lambda_i\right)_{i=1}^n\in\left(0;1\right)$ with $\sum_{i=1}^n\lambda_i=1$ such that

$$\mu^a_{(f;T)}=\sum_{i=1}^n\lambda_i\mu^a_{(f_i;T)}$$

and for every $\hat{T}$ such that $\left(f_i\right)_{i=1}^n$ and $f\in F^{\hat{T}}$,

$$\mu^a_{\left(f;\hat{T}\right)}=\sum_{i=1}^n\lambda_i\mu^a_{\left(f_i;\hat{T}\right)}.$$

It follows that

$$H_a(f;T) = \sum_{i=1}^{n} \lambda_i H_a(f_i;T)$$

for all T such that $(f_i)_{i=1}^n$ and $f \in F^T$. In particular, the coefficients λ_i depend only on the frequencies, but not on the length of the data-sets.

Note that for each $(a';r') \in C$, the sequence of sets $H_a(a';r')^T$ has a limit. In particular, for $a' = a$

$$\lim_{T\to\infty} H_a(a;r')^T = \left\{h \in \Delta^{|R|-1} \mid h \cdot u = u(r')\right\}.$$

For $a \neq a'$, we have, by Axiom 10, that the sequence $\mu^a_{(a';r')^T}$ is either constant or strictly monotonic. Since the sequence is also bounded between $[0;1]$, it follows that the limit

$$\lim_{T\to\infty} \mu^a_{(a';r')^T}$$

exists and we have that:

$$\lim_{T\to\infty} H_a(a';r')^T = \left\{h \in \Delta^{|R|-1} \mid h \cdot u = \lim_{T\to\infty} \mu^a_{(a';r')^T}\right\}.$$

Finally, observe that the limit sets $H_a(a';r')$ are hyperplanes in $\Delta^{|R|-1}$. For the case of $|R| = 3$, they are lines and our non-triviality condition ensures that no two of these lines coincide. For the case $|R| > 3$, they are hyperplanes of dimension 2 or higher.

To summarize, for a given a, the correspondence $H_a(D) : \mathbb{D} \to \Delta^{|R|-1}$ satisfies the following conditions:

1. $H_a(D)$ depends only on the frequency and the length of D, but not on the order of cases in D.
2. Let $(f_i)_{i=1}^n$ and f be in F^T. Whenever $\sum_{i=1}^n \eta_i f_i = f$, for some coefficients $\eta_i \in (0;1)$ such that $\sum_{i=1}^n \eta_i = 1$, there are coefficients $(\lambda_i)_{i=1}^n \in (0;1)$ with $\sum_{i=1}^n \lambda_i = 1$ such that $H_a(f;T) = \sum_{i=1}^n \lambda_i H_a(f_i;T)$.
3. Under the conditions listed in 2., $H_a(f;T) = \sum_{i=1}^n \lambda_i H_a(f_i;T)$ if and only if $H_a\left(f;\hat{T}\right) = \sum_{i=1}^n \lambda_i H_a\left(f_i;\hat{T}\right)$holds for any $\hat{T} \in \mathbb{N}$, such that $(f_i)_{i=1}^n$ and $f \in F^{\hat{T}}$, i.e., λ does not depend on T.
4. The sequences $\left(H_a\left(c^T\right)\right)_{T\in\mathbb{N}}$ have a limit for each $c \in C$.
5. No three of the sets $\lim_{T\to\infty} H_a\left(c^T\right)$ of dimension 0 or 1 are collinear.

As shown in Eichberger and Guerdjikova (2009), conditions 1–5 imply that the sets $H_a(D)$ can

be represented as

$$H_a(D) = \left\{ \frac{\sum_{c \in C} \hat{p}^c_{T,a} s_a(c) f_D(c)}{\sum_{c \in C} s_a(c) f_D(c)} \mid \hat{p}^c_{T,a} \in \hat{P}^c_{T,a} \right\},$$

where $\hat{P}^c_{T.a} = H_a\left(c^T\right)$ and $s_a(c)$ are positive numbers, which are unique up to a multiplication by a positive factor. Furthermore, the definition of $H_a(D)$ implies that

$$V(a; D) = \mu^a_D = u \cdot h \text{ for any } h \in H(a; D),$$

and thus:

$$V(a; D) = \sum u(r) \frac{\sum_{c \in C} \hat{p}^c_{T,a}(r) s_a(c) f_D(c)}{\sum_{c \in C} s_a(c) f_D(c)} \text{ for some } \hat{p}^c_{T,a} \in \hat{P}^c_{T,a}.$$

To see that for $c = (a'; r')$, $s_a(c)$ is independent of r', note that as shown above, for any $D \in \mathbb{D}^T_{a'}$,

$$V(a; D) = \sum_{r \in R} u(r) \hat{p}^c_{T,a}(r) f_D(a'; r)$$

for some $\hat{p}^c_{T,a} \in \hat{P}^c_{T,a}$. It follows that

$$s_a(a'; r) = s_a(a'; r')$$

for all r, $r' \in R$ and the functions $s_a(c)$ can be equivalently written as $s_a(a_c)$, where a_c is the action chosen in case c.

It remains to derive the coefficients α and β and the sequence γ_T and show that they have the properties stated in the Theorem. Define α_T, β_T and γ_T by:

$$\begin{aligned} \beta_T &= : 1 - \mu^a_{(a;\bar{r})^T} \\ \alpha_T &= : \mu^a_{(a;\underline{r})^T} \\ \gamma_T &= : \alpha_T + \beta_T. \end{aligned}$$

Observe that

$$\lim_{T \to \infty} \alpha_T = \lim_{T \to \infty} \beta_T = \lim_{T \to \infty} \gamma_T = 0.$$

We now wish to show that for each $D \in \mathbb{D}^T_a$,

$$\begin{aligned} V(a; D) &= \alpha_T u(1) + \beta_T u(0) + (1 - \gamma_T) \sum_{r \in R} u(r) f_D(a; r) \\ &= \alpha_T + (1 - \alpha_T - \beta_T) \sum_{r \in R} u(r) f_D(a; r). \end{aligned}$$

Let first $D = (a;r)^T$. Then,

$$\begin{aligned} V\left(a;(a;r)^T\right) &= \mu^a_{(a;r)^T} = \lambda_r \mu^a_{(a;\bar{r})^T} + (1-\lambda_r)\,\mu^a_{(a;\underline{r})^T} \\ &= \lambda_r\,(1-\beta_T) + (1-\lambda_r)\,\alpha_T \\ &= \alpha_T + \lambda_r\,(1-\alpha_T-\beta_T) \\ &= \alpha_T + u\,(r)\,(1-\alpha_T-\beta_T)\,. \end{aligned}$$

More generally, for $D \in \mathbb{D}^T_a$, we have that

$$\begin{aligned} V\,(a;D) &= \mu^a_D = \sum_{r\in R} f_D\,(a;r)\,\mu^a_{(a;r)^T} \\ &= \alpha_T + (1-\alpha_T-\beta_T)\sum_{r\in R} f_D\,(a;r)\,u\,(r) \\ &= \alpha_T u\,(1) + \beta_T u\,(0) + (1-\gamma_T)\sum_{r\in R} u\,(r)\,f_D\,(a;r)\,. \end{aligned}$$

It immediately follows that for all $D \in \mathbb{D}^T_a$,

$$H_a\,(D) = \left\{h \in \Delta^{|R|-1} \mid h\cdot u = \left[\alpha_T\delta_{(a;\bar{r})} + \beta_T\delta_{(a;\underline{r})} + (1-\gamma_T)\,f_D\,(a;r)\right]\cdot u\right\}$$

and, hence,

$$\lim_{T\to\infty} \hat{P}^{(a;r)}_{a,T} = \lim_{T\to\infty} H_a\,(a;r)^T = \left\{h \in \Delta^{|R|-1} \mid h\cdot u = u\,(r)\right\}.$$

For $D = (a';r')^T$ with $a' \neq a$, we know by Axiom 6 that there is a $\lambda_{(a';r')} \in [0;1]$ such that

$$\begin{aligned} V\left(a;(a';r')^T\right) &= \mu^a_{(a';r')^T} = \lambda_{(a';r')}\mu^a_{(a;\bar{r})^T} + \left(1-\lambda_{(a';r')}\right)\mu^a_{(a;\underline{r})^T} \\ &= \lambda_{(a';r')}\,(1-\beta_T) + \left(1-\lambda_{(a';r')}\right)\alpha_T \\ &= \alpha_T + \lambda_{(a';r')}\,(1-\alpha_T-\beta_T) \\ &= \alpha_T + \lambda_{(a';r')}\,(1-\alpha_T-\beta_T)\,. \end{aligned}$$

By Axiom 8, this $\lambda_{(a';r')}$ is independent of T, hence, we can define $h^a_{(a';r')}$ to be such that:

$$u\cdot h^a_{(a';r')} = \lambda_{(a';r')}.$$

Since $\lambda_{(a';r')} \in [0;1]$, such an $h^a_{(a';r')}$ exists. Hence,

$$V\left(a;(a';r')^T\right) = \alpha_T u\,(1) + (1-\gamma_T)\sum_{r\in R} u\,(r)\,h^a_{(a';r')}\,(r) + \beta_T u\,(0)\,.$$

It follows that

$$H_a\left(a';r'\right)^T=\left\{h\in\Delta^{|R|-1}\mid h\cdot u=\left[\alpha_T\delta_{(a;\bar{r})}+\beta_T\delta_{(a;\underline{r})}+\left(1-\gamma_T\right)h^a_{(a';r')}\right]\cdot u\right\} \tag{21}$$

We now wish to show that the ratio $\frac{\alpha_T}{\alpha_T+\beta_T}$ does not depend on T. By Axiom 9, there is an outcome $\hat{r}$ such that:

$$\left(a;\left(a;\hat{r}\right)^k\right)\sim\left(a;\left(a;\hat{r}\right)\right)$$

for all $k\in\mathbb{N}$. This is equivalent to:

$$\begin{aligned} V\left(a;\left(a;\hat{r}\right)^T\right) &= \\ \alpha_T+\left(1-\alpha_T-\beta_T\right)u\left(\hat{r}\right) &= u\left(\hat{r}\right)\mu^a_{(a;\bar{r})^T}+\left(1-u\left(\hat{r}\right)\right)\mu^a_{(a;\underline{r})^T} \end{aligned}$$

for all $T\in\mathbb{N}$. In particular, since $\lim\mu^a_{(a;\bar{r})^T}=1$ and $\lim\mu^a_{(a;\underline{r})^T}=0$, it must be that:

$$V\left(a;\left(a;\hat{r}\right)^T\right)=u\left(\hat{r}\right)$$

for all $T\in\mathbb{N}$. It follows that:

$$\begin{aligned} \alpha_T+\left(1-\alpha_T-\beta_T\right)u\left(\hat{r}\right) &= u\left(\hat{r}\right) \\ \alpha_T-\left(\alpha_T+\beta_T\right)u\left(\hat{r}\right) &= 0 \\ \frac{\alpha_T}{\left(\alpha_T+\beta_T\right)} &= u\left(\hat{r}\right). \end{aligned} \tag{22}$$

Denote $\alpha=:\frac{\alpha_T}{(\alpha_T+\beta_T)}$ *and* $\beta=:\frac{\beta_T}{(\alpha_T+\beta_T)}$*. Then, (*21*) implies that for any* $(a';r')$*,*

$$H_a\left(a';r'\right)^T=\left\{h\in\Delta^{|R|-1}\mid h\cdot u=\left[\gamma_T\left(\alpha\delta_{(a;\bar{r})}+\beta\delta_{(a;\underline{r})}\right)+\left(1-\gamma_T\right)h^a_{(a';r')}\right]\cdot u\right\}$$

It remains to show that $\gamma_T\in(0;1)$ *for all* T*. Note that for any* $r\in R$*, we can write the utility of* $(a;r)^T$ *as:*

$$V\left(a;r\right)=\gamma_T\alpha+\left(1-\gamma_T\right)u\left(r\right)=\gamma_Tu\left(\hat{r}\right)+\left(1-\gamma_T\right)u\left(r\right).$$

Suppose first that there is a case $(a;r)$ *such that* $(a;r)\succ(a;\hat{r})$*, then we have by Axiom 10 that* $(a;r)^{T+1}\succ(a;r)^T$ *for all* T*. Since* $\lim_{T\to\infty}V\left(a;r\right)^T=u\left(r\right)$ *and since* $V\left(a;r\right)^T$ *is a strictly increasing sequence, the first element of which* $V\left(a;r\right)>u\left(\hat{r}\right)$*, we have that:*

$$V\left(a;r\right)^T\in\left(u\left(\hat{r}\right);u\left(r\right)\right)$$

for all T*. It follows that* $\gamma_T\in(0;1)$ *for all* T*.*

Suppose now that there is no $(a;r)$ *satisfying* $(a;r)\succ(a;\hat{r})$*, but there exists an* $(a;r)$ *such that*

$(a;\hat{r}) \succ (a;r)$. *Then, by Axiom 10,* $(a;r)^T \succ (a;r)^{T+1}$ *for all* T. *Since* $\lim_{T\to\infty} V(a;r)^T = u(r)$ *and since* $V(a;r)^T$ *is a strictly decreasing sequence, the first element of which* $V(a;r) < u(\hat{r})$*, we have that:*

$$V(a;r)^T \in (u(r);u(\hat{r}))$$

for all T*, which again implies that* $\gamma_T \in (0;1)$. *Note that the case in which* $(a;r) \sim (a;\hat{r})$ *is excluded by the non-triviality condition.*

As we showed above, γ_T *does not depend on the data-set to be evaluated, hence, this concludes the proof of the Theorem.*■

9 References

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

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

BEWLEY, T. F. (1986). "Knightian Decision Theory: Part I", Discussion Paper, *Cowles Foundation*.

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

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

EASLEY,. D., RUSTICHINI, A. (1999). "Choice Without Beliefs", *Econometrica 67,* 1157-1184.

EICHBERGER, J., AND GUERDJIKOVA, A. (2009). "Case-Based Belief Formation under Ambiguity", *mimeo,* Cornell University.

ELLSBERG, D. (1961). "Risk, Ambiguity and the Savage Axioms", *Quarterly Journal of Economics 75*, 643-669.

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

FISHBURN, P. (1970). *Utility Theory for Decision Making* , New York: John Wiley and Sons.

GAJDOS, TH., HAYASHI, T., TALLON, AND J.-M., VERGNAUD, J.-C. (2007). "Attitude Towards Imprecise Information", *Journal of Economic Theory, 141,* 68-99.

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

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

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

UK: Cambridge University Press.

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

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

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

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

GRANT, S., KAJI, A., POLAK, B. (1998). "Instrinsic Preference for Information", *Journal of Economic Theory, 83,* 233-259.

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

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

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

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", *Working paper,* University of Texas, Austin.