

Decision making on the sole basis of statistical likelihood

Phan H. Giang^{a,*}, Prakash P. Shenoy^b

^a *Computer Aided Diagnosis and Therapy Solutions (CAD), Siemens Medical Solutions,
51 Valley Stream Pkwy, Malvern, PA 19355, USA*

^b *University of Kansas School of Business, 1300 Sunnyside Ave, Summerfield Hall,
Lawrence, KS 66045-7585, USA*

Received 10 May 2004; accepted 16 March 2005

Available online 10 May 2005

Abstract

This paper presents a new axiomatic decision theory for choice under uncertainty. Unlike Bayesian decision theory where uncertainty is represented by a probability function, in our theory, uncertainty is given in the form of a likelihood function extracted from statistical evidence. The likelihood principle in statistics stipulates that likelihood functions encode all relevant information obtainable from experimental data. In particular, we do not assume any knowledge of prior probabilities. Consequently, a Bayesian conversion of likelihoods to posterior probabilities is not possible in our setting. We make an assumption that defines the likelihood of a set of hypotheses as the maximum likelihood over the elements of the set. We justify an axiomatic system similar to that used by von Neumann and Morgenstern for choice under risk. Our main result is a representation theorem using the new concept of binary utility. We also discuss how ambiguity attitudes are handled. Applied to the statistical inference problem, our theory suggests a novel solution. The results in this paper could be useful for probabilistic model selection.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Decision theory; Likelihood; Statistical inference; Ambiguity attitude

* Corresponding author.

E-mail addresses: phan.giang@siemens.com (P.H. Giang), pshenoy@ku.edu (P.P. Shenoy).

1. Introduction

Various formal decision theories for choice under risk and uncertainty have been studied since the seminal work by von Neumann and Morgenstern (vNM) [38] where the expected utility maximization principle was formally established. With few exceptions, a common feature in these theories is the use of probability to express uncertainty in decision situations. An axiomatic model is as good as its axioms, the debate on axioms of the vNM theory started almost immediately with their publication [2]. As a result of this ongoing debate, axiomatic systems that are weaker than vNM but still possess the expected utility representation have been investigated [17,32,33]. There is also a recognition that the uncertainty that one usually associates with the words “ambiguity”, “vagueness” and “fuzziness” are not the same kind as that associated with “risk”. The latter is captured by standard numerical probability.

In this paper,¹ we consider a class of choice problems where uncertainty is characterized by likelihood functions. This class includes a typical statistical inference problem that is formulated as follows. Suppose we are to analyze a statistical experiment on a random variable Y given (i) Y follows one of the distributions in $\mathcal{F} = \{P_\theta \mid \theta \in \Omega\}$ parameterized by θ ; and (ii) the outcome of the experiment is $Y = y$. The question is: what can we conclude about the true value of parameter θ ?

There is consensus among statisticians about what information sample y brings to the unknown parameter. According to the likelihood principle, one of the fundamental principles of statistics [4,5,8], all relevant information of the sample is encoded in the likelihood function on the parameter space. And the consensus also ends at this point. The statistical inference problem is treated differently by different approaches [3].

According to the decision-theoretic approach advocated by Wald [39], the inference problem is viewed as a choice problem. For example, in the context of a hypothesis testing problem, the choice is to either accept or reject a hypothesis. Within the decision-theoretic approach there are several variations. Wald’s *maximin* decision rule selects an action that delivers the most favorable worst-case outcome. A Bayesian treatment of the problem suggests a calculation of posterior probability function on Ω via Bayes’s theorem from the likelihood function by assuming a prior distribution. Given the posteriors, actions are compared on the basis of their expected utility. In this paper, we propose a third alternative. We construct a decision theory that works directly with likelihood information. We choose to treat likelihood as uncertainty in its own right for a simple reason: priors are not known in many situations.

The problem of probabilistic model selection in the areas of AI, machine learning, pattern recognition and data mining is an example of the statistical inference problem. Given a (training) data set y , researchers construct a probabilistic model P (e.g., a Bayesian net) that generates/fits the data and then use this model for inference with future observations. Because there are, almost always, more than one models that emerge as plausible candidates, model selection is an essential part of model construction.

¹ A preliminary version of this work has appeared in the Proceedings of 18th Conference on Uncertainty in Artificial Intelligence (UAI 2002) [21].

From a decision theoretic point of view, the prevailing practice to select a model using simple criteria such as maximum likelihood (ML) or maximum a posteriori probability (MAP) is equivalent to assuming equal utilities (costs) for all models under consideration. This assumption however does not explain the following example. Given two models of approximately the same likelihoods, most researchers would go for a simpler one (and justify this choice by invoking Occam’s razor principle). In the statistics literature, models are selected by using Akaike Information Criterion (AIC) [1] or Schwarz’s criterion (a.k.a. Bayesian Information Criterion or BIC) [34]. The idea underlying both AIC and BIC is to penalize model’s likelihood by an amount depending on its number of parameters. Poland and Shachter [31] suggest the “effectiveness ratio” criterion where the penalty has an explicit computational interpretation. Clearly, the concern on complexity can be viewed as a cost associated with a model. In broader terms, an implication from these works is that different models are associated with *different* costs that must be taken into account in a model selection process.

This paper is organized as follows. In the next section, we discuss extending a likelihood function to a function on the set of subsets of possible worlds. In the main part (Section 3), we develop a decision theory for likelihood uncertainty. We begin by proposing a set of five axioms that are justified by intuition as well as by the stochastic dominance principle. Next, we introduce the concept of binary utility and prove the representation theorem for likelihood lotteries. That is followed by comments on related works. In Section 4, our decision theory is applied for a statistical inference problem. Section 5 contains some concluding remarks.

2. Likelihood as uncertainty measure

Let us consider the statistical inference problem as described earlier. Although the phenomenon under study is described probabilistically (by a set of probability functions \mathcal{F}), the uncertainty pertaining to the choice problem is not. It is a likelihood function. The term ‘likelihood’ used in modern statistics was coined by R.A. Fisher who mentioned it as early as 1922 [18]. Fisher used likelihoods to measure “mental confidence” in competing scientific hypotheses as a result of a statistical experiment (see [14] for a detailed account). Likelihood has a puzzling nature. For each $\theta \in \Omega$, there is a likelihood quantity that by magnitude equals $P_\theta(y)$ —the probability (or probability density in case of infinite Ω)² of observing y if θ is in fact the true value of the parameter. However, if we view the set of likelihood quantities as a function on the parameter space, we have a likelihood function. A likelihood function is not a probability function. For a simple reason, the sum of all likelihood values (over the parameter space) may not add to unity. Moreover, likelihood functions are equivalent up to a proportional constant.

To emphasize the fact that a likelihood function is tied to data y and has θ as the variable, the notation $lik_y(\theta)$ is used instead of $P_\theta(y)$. Technically, probability and likelihood are

² One can write $P_{\theta_0}(y)$ in the form of a conditional probability: $P(Y = y \mid \theta = \theta_0)$. The latter notation implies that there is a probability measure on parameter space Ω . This is the case for the Bayesian approach. In this paper, we do not assume such a probability measure. So we will stick with the former notation.

two kinds of animals, but they are as close as mule and donkey. This proximity is the reason for an intertwining relationship. Obviously, (posterior) probability is derived from likelihood and priors via Bayes theorem. Since such priors are supposed to summarize the information about the parameter *before* the experiment is conducted, the assumption of its existence is beyond the realm of science as many statisticians contend. Although in certain situations prior probability comes naturally, there is no compelling argument why it must *always* be known to an experimenter in *all* situations.

Another path from likelihood to probability, that bypasses the issue of priors, was started by Fisher himself. He suggested to compute what he called *fiducial* probabilities by normalizing the likelihoods (dividing by the sum of likelihoods). Fisher's idea has been shown to work for isolated examples and but it faces a serious difficulty when applied to general cases. Some statisticians now believe that the fiducial probability is a mistake [3].

Belief function theory was proposed by Dempster [9] in an attempt to overcome the difficulty of the fiducial argument. Shafer [35] is mainly responsible for turning Dempster's idea into a full-fledged theory of evidence. A basic construct in Dempster–Shafer theory is *basic probability assignment* (BPA)

$$m: 2^\Omega \rightarrow [0, 1] \quad \text{such that } m(\emptyset) = 0 \text{ and } \sum_{A \subseteq \Omega} m(A) = 1 \quad (1)$$

Value $m(A)$ for $A \subseteq \Omega$ is called probability *mass* of A . If $m(A) > 0$ then A is called a *focus*. A standard probability function is a belief function whose foci are singletons. From a BPA, one can derive a *plausibility* function

$$Pl(A) \stackrel{\text{def}}{=} \sum_{B \cap A \neq \emptyset} m(B) \quad \text{for all } \emptyset \neq A \subseteq \Omega \quad (2)$$

BPA and plausibility have the same information content since the original m can be recovered from Pl .

Shafer [35] proposes to represent statistical evidence by a belief function with nested foci (*consonant belief function* or *CBF*) such that plausibilities on singletons are proportional to the likelihood values. Given a likelihood function lik_x , the corresponding CBF is constructed as follows. Suppose lik_x partitions Ω into $\{\Omega_i\}$ according to its values $\Omega_i = \{\omega \mid lik_x(\omega) = a_i\}$ with $a_1 > a_2 > \dots > a_k$. Then there are k foci (A_i) and k masses

$$A_i = \bigcup_{j=1}^i \Omega_j \quad \text{and} \quad m(A_i) = \frac{(a_i - a_{i+1})}{a_1} \quad (3)$$

where $a_{k+1} \stackrel{\text{def}}{=} 0$.

A consequence of nested-focus structure is that plausibility function is union decomposable, i.e., for $A, B \subseteq \Omega$

$$Pl(A \cup B) = \max(Pl(A), Pl(B)) \quad (4)$$

A crucial argument in favor of Shafer's original³ proposal is that likelihood treatment in CBF is in agreement with the maximum likelihood method (ML) of statistics. In ML, the likelihood assigned to a set of hypotheses is taken to be the maximum of the likelihoods of individual hypothesis in the set. The idea of taking the maximum individual likelihood as the likelihood for a set has been a standard practice since the publication of seminal papers [30] by Neyman and Pearson (1928). ML is not only intuitively appealing, but it is also backed by various asymptotic optimality properties [26,27].

We will use different notation than one in [35]. We want to emphasize the nature of likelihood and avoid belief function connotations. While Shafer is mainly interested in representing and reasoning with evidence, our goal is decision making. Let us define an *extended likelihood function* or *ELF* $Lik_y : 2^\Omega \rightarrow [0, 1]$ as follows.

$$Lik_y(\theta) \stackrel{\text{def}}{=} \frac{lik_y(\theta)}{\sup_{\omega \in \Omega} lik_y(\omega)} = \frac{lik_y(\theta)}{lik_y(\hat{\theta})} \quad \text{for } \theta \in \Omega \quad (5)$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ .

$$Lik_y(A) \stackrel{\text{def}}{=} \sup_{\omega \in A} Lik_y(\omega) \quad \text{for } A \subseteq \Omega \quad (6)$$

After learning that the true value of parameter is in a subset of the parameter space $B \subseteq \Omega$ such that $Lik_y(B) > 0$, one should *condition* the ELF by the following equation.

$$Lik_y(A | B) \stackrel{\text{def}}{=} \frac{Lik_y(A \cap B)}{Lik_y(B)} \quad (7)$$

This definition of likelihood conditioning is derived from Dempster's rule of combination applied for a consonant belief function in plausibility form. It also conforms to the likelihood principle as the following example demonstrates.

We use the convention $Lik_y(\emptyset) = 0$. Some properties of Lik follow directly from its definitions.

Lemma 1.

- (i) $Lik_y(\Omega) = 1$.
- (ii) $Lik_y(A \cup B) = \max\{Lik_y(A), Lik_y(B)\}$.
- (iii) $\max\{Lik_y(A), Lik_y(\bar{A})\} = 1$ where \bar{A} is the complement of A in Ω .
- (iv) If $A \subseteq B$ then $Lik_y(A) \leq Lik_y(B)$.

Example. A r.v. Y is known to have a normal distribution. It is also known that mean $\mu \in \{0, 1\}$ and standard deviation $\sigma \in \{1, 1.5\}$. Suppose that value $y = 1.4$ is observed. We

³ Shafer [36] later renounces his idea on the ground that the set of CBFs is not closed under Dempster's rule of combination, which is the standard rule to combine two distinct pieces of evidence. It implies that representation of compound evidence is not the same as the combination of individual evidences. However, later Walley [40] shows that Dempster's rule is not compatible with the likelihood principle, and therefore is not suitable for combining statistical evidence. He also shows that set of CBFs is closed under an alternative combination rule. With respect to conditioning Dempster's rule and Walley's alternative are identical.

Table 1
Likelihood, extended likelihood and conditioning

	ω_1	ω_2	ω_3	ω_4	A	B	C	D
$lik_{1.4}(\cdot)$	0.1497	0.1721	0.3683	0.2567	n.a.	n.a.	n.a.	n.a.
$Lik_{1.4}(\cdot)$	0.4065	0.4673	1.0000	0.6970	0.4673	1	1	0.6970
$Lik_{1.4}(\cdot D)$	0.0000	0.6704	0.0000	1.0000	0.6704	1	0	1.0000

want to calculate the ELF representing uncertainty about unknown parameters. The unknown parameter $\theta = (\mu, \sigma)$. $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ with $\omega_1 = (0, 1)$, $\omega_2 = (0, 1.5)$, $\omega_3 = (1, 1)$, $\omega_4 = (1, 1.5)$. Let A denote event $\mu = 0$, B denote $\mu = 1$, C denote $\sigma = 1$ and D denote $\sigma = 1.5$. That means $A = \{\omega_1, \omega_2\}$, $B = \{\omega_3, \omega_4\}$, $C = \{\omega_1, \omega_3\}$ and $D = \{\omega_2, \omega_4\}$. In the first row are densities at 1.4 of normal probability density function given a configuration of mean μ and standard deviation σ : $f(1.4 | \mu, \sigma)$. For example, $f(1.4 | \mu = 0, \sigma = 1) = 0.1497$. Obviously, density is not defined for a set of configurations (A , B , C or D). Eq. (6) is used to calculate the second row.

Suppose in addition to that, it becomes known that $\sigma = 1.5$. From a statistical point of view, in the new situation mean μ is the unknown parameter of interest. The likelihoods of $\mu = 0$ and $\mu = 1$ are 0.1721 and 0.2567 respectively. The likelihood ratio is 0.6704. In terms of extended likelihood, the new situation is coded by conditional $Lik_{1.4}(\cdot | D)$. This yields $Lik_{1.4}(\omega_1 | D) = Lik_{1.4}(\omega_3 | D) = Lik_{1.4}(C | D) = 0$ and $Lik_{1.4}(A | D) = Lik_{1.4}(\omega_2)/Lik_{1.4}(D) = 0.6704$ and $Lik_{1.4}(B | D) = Lik_{1.4}(\omega_4)/Lik_{1.4} = 1$. The ratio of extended likelihoods of $\mu = 0$ and of $\mu = 1$ is 0.6704. Keeping in mind that the likelihood principle holds that the ratio of likelihoods is what matters, we see that our definition of conditioning of extended likelihoods conforms to the standard practice in statistics.

It is worth noting that while our derivation of the extended likelihood (CBF) is motivated by statistical considerations, the properties listed in Lemma 1 are also the defining properties of a possibility measure [12,41]. What distinguishes a possibility function from a CBF is its fuzzy set semantics and, consequently, the notion of *ordinal conditioning*:

$$\pi(A | B) = \begin{cases} 1 & \text{if } \pi(A \cap B) = \pi(A) \\ \pi(A \cap B) & \text{otherwise} \end{cases} \quad (8)$$

Technically, possibility theory can entertain both notions of conditioning: the ordinal (Eq. (8)) and the *numerical* one specified by Eq. (7). In this sense, extended likelihood is a possibility measure equipped with numerical conditioning.⁴ It is not difficult to check (using the previous example) that if Lik was updated using ordinal conditioning then the result would not be consistent with the likelihood principle.

⁴ Dubois, Moral and Prade [11] show that a possibility measure is the result of taking supremum on a family of likelihood functions. On that semantics, the min rule for combination of possibility measures is justified.

3. A decision theory with likelihood uncertainty

Let us formalize the decision problem we will study. We assume a decision situation described by a tuple $(\Omega, X, \mathcal{A}, y, \pi)$. Ω is the set of simple hypotheses (the parameter space); X is the set of rewards measured in a utility currency; \mathcal{A} is a set of actions of the form $\Omega \rightarrow X$; y is the data/observation/evidence and π is the ELF $2^\Omega \rightarrow [0, 1]$ calculated from y representing uncertainty on hypotheses.

For actions in this setting, the states on which rewards are contracted are not observable. The utility nature of rewards implies that the risk attitude factor in decision making can be ignored. Also, since utility can be transformed by a linear function, we can assume that rewards are restricted to the unit interval $[0, 1]$. For the sake of clarity, we assume that X is finite and its elements are denoted by x_1, x_2, \dots, x_r .

A *simple* likelihood lottery is an action coupled with a likelihood measure. Each lottery is a mechanism that delivers rewards with associated likelihoods. Formally, a lottery L induced by π and a is a mapping from $X \rightarrow [0, 1]$ such that $L(x) = \pi(a^{-1}(x))$ for $x \in X$ where a^{-1} is a set-valued inverse mapping of action a . For the remainder of this paper, we denote a simple lottery by $[L(x_1)/x_1, L(x_2)/x_2, \dots]$ with convention that those x_j for which $L(x_j) = 0$ are omitted. In this notation, a consequence $x \in X$ is identified with a unary lottery $[1/x]$. Notice that for any lottery $[L_i/x_i]_{i=1}^m$, $\bigcup_{1 \leq i \leq m} a^{-1}(x_i) = \Omega$. Since π is an ELF and $L_i = \pi(a^{-1}(x_i))$, therefore, $\max_{1 \leq i \leq m} L_i = 1$.

We also consider *compound* lotteries whose rewards are other lotteries. The set of lotteries is denoted by \mathcal{L} .

3.1. Axioms

We study preference relation \succeq on the set of lotteries \mathcal{L} ($\succeq \subseteq \mathcal{L}^2$). Indifference \sim and strict preference \succ relations are derived from \succeq . $L_1 \sim L_2$ iff $L_1 \succeq L_2$ & $L_2 \succeq L_1$. $L_1 \succ L_2$ iff $L_1 \succeq L_2$ & $L_2 \not\succeq L_1$. We postulate that \succeq satisfies five axioms similar to those proposed by von Neumann and Morgenstern for the classical linear utility theory (in the form presented in [29]). They are as follows.

A1 Order. \succeq is reflexive, transitive and complete.

Since the consequences in X are special lotteries, \succeq is also the order on consequences. We can assume that $x_1 \succeq x_2 \succeq \dots \succeq x_r$ with $x_1 \succ x_r$. In some cases to make clear we are dealing with the best and the worst consequences, special notations are used for x_1 and x_r namely, $\bar{x} \equiv x_1$ and $\underline{x} \equiv x_r$. A lottery that involves only the best \bar{x} and the worst consequences \underline{x} as potential outcomes is called a *canonical* lottery. The set of canonical lotteries is denoted by \mathcal{L}_c .

A2 Reduction of compound lotteries.

Let $L = [\delta_1/L_1, \delta_2/L_2, \dots, \delta_k/L_k]$ and $L_i = [\kappa_{i1}/x_1, \kappa_{i2}/x_2, \dots, \kappa_{ir}/x_r]$ then $L \sim [\kappa_1/x_1, \kappa_2/x_2, \dots, \kappa_r/x_r]$ with $\kappa_j = \max_{1 \leq i \leq k} \{\delta_i \cdot \kappa_{ij}\}$.

A3 Substitutability.

If $L_i \sim L'_i$ then $[\delta_1/L_1, \dots, \delta_i/L_i \dots \delta_k/L_k] \sim [\delta_1/L_1, \dots, \delta_i/L'_i \dots \delta_k/L_k]$.

A4 Existence of equivalent canonical lottery.

For each $x \in X$ there is a $s \in \mathcal{L}_c$ such that $x \sim s$.

A5 Qualitative monotonicity.

$$[\lambda/\bar{x}, \mu/\underline{x}] \succeq [\lambda'/\bar{x}, \mu'/\underline{x}] \quad \text{iff} \quad (\lambda \geq \lambda') \& (\mu \leq \mu') \quad (9)$$

Among the axioms, A1 and A3 are standard assumptions about a preference relation.

A2 is an implication of the conditioning operation. Suppose that the unknown parameter θ is a vector. We can think, for example, $\theta = (\gamma, \sigma)$. Let us consider a compound lottery $L = [\delta_1/L_1, \delta_2/L_2, \dots, \delta_k/L_k]$ where $L_i = [\kappa_{i1}/x_1, \dots, \kappa_{ir}/x_r]$ for $1 \leq i \leq k$. Underlying L , in fact, is a two-stage lottery. The first stage is associated with a scalar parameter γ . It accepts values $\gamma_1, \gamma_2, \dots, \gamma_k$ with likelihoods $\delta_1, \delta_2, \dots, \delta_k$ respectively. If γ_i is the true value, the holder of L is rewarded with simple lottery L_i that, in turn, is associated with scalar parameter σ that accepts $\sigma_{o_i(1)}, \sigma_{o_i(2)}, \dots, \sigma_{o_i(r)}$ with likelihoods $\kappa_{i1}, \kappa_{i2}, \dots, \kappa_{ir}$ where o_i is a permutation of $(1, 2, 3, \dots, r)$. When $\sigma_{o_i(j)}$ obtains, the holder is rewarded with consequence x_j .

Let us consider another one-stage lottery L' that delivers x_j in case tuple $\langle \gamma_i \sigma_{o_i(j)} \rangle$ is the true value of θ for $1 \leq i \leq k$. Because of conditioning equation (7), we have

$$Lik(\gamma_i \sigma_{o_i(j)}) = Lik(\gamma_i) Lik(\sigma = \sigma_{o_i(j)} \mid \gamma = \gamma_i) = \delta_i \cdot \kappa_{ij} \quad (10)$$

The set of tuples for which x_j is delivered is $\{\langle \gamma_i \sigma_{o_i(j)} \rangle \mid 1 \leq i \leq k\}$. Thus, the extended likelihood associated with consequence x_j in lottery L'

$$Lik(\{\gamma_i \sigma_{o_i(j)} \mid 1 \leq i \leq k\}) = \max\{\delta_i \cdot \kappa_{ij} \mid 1 \leq i \leq k\} \quad (11)$$

Since L and L' have the property that no matter what is the true value of θ , the consequences they deliver are always the same, we require $L \sim L'$ which is axiom A2. Fig. 1 shows an example where $k = 2$ and $r = 2$, $o_1(1) = 1$, $o_1(2) = 2$, $o_2(1) = 2$ and $o_2(2) = 1$.

Axiom A4 requires that for any consequence $x \in X$ there is a canonical lottery $c = [\lambda_1/\bar{x}, \lambda_2/\underline{x}]$ such that $x \sim c$. For clarity, let us assume that $\bar{x} = 1$, $\underline{x} = 0$ (the argument remains valid for any real values of \bar{x} and \underline{x} as long as $\bar{x} > \underline{x}$). For any $x \in X$, we need to find a canonical lottery c equivalent to x . We will describe a *likelihood gamble* for this purpose.

There are three parties in this game: the Arbiter, the House and the Player. The goal of the game is to gauge binary utility function for the Player. The game plays as follows. The Arbiter preselects a single parameter probability distribution f_θ . She also predetermines

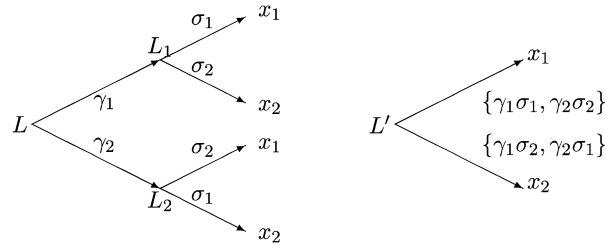


Fig. 1. Two-stage L and one-stage L' are equivalent.

2 values $\{\theta_1, \theta_2\}$ that she will use for the unknown parameter. Probability distribution f_θ as well as the possible values of the parameter are revealed to all parties. For example, a normal distribution⁵ with given standard deviation $\sigma = 1$ can be used as f_θ where θ is unknown mean and $\theta \in \{-1, +1\}$.

The Arbiter secretly picks a value θ in $\{\theta_1, \theta_2\}$ then generates a value y from f_θ . The mechanism used by the Arbiter to pick the value θ is unknown to both the Player and the House. Both the Player and the House are told about the data y . The House offers a gamble that pays \bar{x} to the Player if the value actually used by the Arbiter to generate the observation is θ_1 and \underline{x} if it is θ_2 . What is the highest price the Player would be willing to pay for this gamble? If the answer is x , then for the Player,

$$x \sim [Lik_y(\theta_1)/\bar{x}, Lik_y(\theta_2)/\underline{x}] \quad (12)$$

where $Lik_y(\theta_1) \propto f_{\theta_1}(y)$ and $Lik_y(\theta_2) \propto f_{\theta_2}(y)$.

One can repeat this gamble any number of times to get a table of correspondence between x_i and $[Lik_{y_i}(\theta_1)/\bar{x}, Lik_{y_i}(\theta_2)/\underline{x}]$. What we assume in A4 is that we can make the table rich enough so that for any $x \in X$ we can look up the table for an equivalent $[Lik_y(\theta_1)/\bar{x}, Lik_y(\theta_2)/\underline{x}]$.

Comparing the likelihood gamble with a probabilistic gamble⁶ used in practice to extract decision maker's (unary) utility, we see a number of important differences. First, instead of r.v. with a known distribution, e.g., tossing a fair coin or rolling a dice, a partially specified probability model is used. Second, the rewards for the Player in the likelihood gamble are contracted, not on observations (y) but on the unobservable true value of the parameter (θ). In this sense, a likelihood gamble is a simple hypothesis testing problem because the Player needs to decide which of two hypotheses $\theta = \theta_1$ or $\theta = \theta_2$ is true. The relationship is made clear in Fig. 5.

What kind of betting behavior should be expected from a rational decision maker? Certain patterns should be excluded as irrational. In the example of a likelihood gamble that uses a normal distribution with known s.d. $\sigma = 1$ and pays \$1 if mean is -1 and 0 if it is 1, paying \$0.20 for the gamble if $y = -3$ and paying \$0.70 if $y = 1$ would be irrational. Intuitively, $y = -3$ lends more support to hypothesis $\theta = -1$ than $y = 1$ does. Let us formalize this intuition. We impose a mild constraint in the form of monotonicity axiom A5. Basically, we require that the price for lottery $[\lambda/1, \mu/0]$ is greater or equal to the price for $[\lambda'/1, \mu'/0]$ if the likelihood of getting 1 in the former is higher than that of the latter ($\lambda \geq \lambda'$) and likelihood of getting 0 in the former is less than that of the latter ($\mu \leq \mu'$).

We justify A5 on the basis of first order stochastic dominance (FSD). Not being strictly Bayesian, we won't assume to know the prior probability of $P(\theta = \theta_1)$, but we will assume that such a prior exists. This situation can be modeled by viewing the prior of $\theta = \theta_1$ as a r.v. ρ taking value in the unit interval. The distribution of ρ is unknown to us. We calculate the posterior of $\theta = \theta_1$ given y and ρ .

$$P_\rho(\theta = \theta_1 | y) = \frac{\rho \cdot Lik_y(\theta_1)}{\rho \cdot Lik_y(\theta_1) + (1 - \rho) \cdot Lik_y(\theta_2)} \quad (13)$$

⁵ Any other distribution will work as fine.

⁶ Suppose $u(\$0) = 0$ and $u(\$1) = 1$, and d the price a decision maker is willing to pay for a gamble that pays \$1 if a toss of a fair coin turns Head and \$0 if it turns Tail, then $u(d) = 0.5$.

The expected payoff of the likelihood gamble

$$V_y(\rho) = P_\rho(\theta = \theta_1 | y) \cdot \bar{x} + P_\rho(\theta = \theta_2 | y) \cdot \underline{x} \quad (14)$$

$$= \underline{x} + P_\rho(\theta = \theta_1 | y) \cdot (\bar{x} - \underline{x}) \quad (15)$$

With $\bar{x} - \underline{x} > 0$, $V_y(\rho)$ is a strictly increasing function of $P_\rho(\theta = \theta_1 | y)$. When $\bar{x} = 1$ and $\underline{x} = 0$, Eq. (14) is further simplified to $V_y(\rho) = P_\rho(\theta = \theta_1 | y)$. Being a function of ρ , $V_y(\rho)$ is a r.v.

The concept of *stochastic dominance* (SD) has been used extensively in economics, finance, statistics [28]. Suppose X and Y are two distinct r.v. with the cumulative distributions F and G respectively. We say that X stochastically dominates (to the first degree) Y (written as XD_1Y) iff $F(x) \leq G(x) \forall x$. Since X and Y are distinct, strict inequality must hold for at least one value x . FSD is important because of the following equivalence: X stochastically dominates (first order) Y iff the expected utility of X is greater than or equal to the expected utility of Y for all non-decreasing utility functions i.e., XD_1Y iff $E(u(X)) \geq E(u(Y)) \forall u \in U$ where U the class of non-decreasing utility functions and $E(\cdot)$ is the expectation operator.

In the immediately following discussion, we will assume $\bar{x} = 1$ and $\underline{x} = 0$ for the sake of clarity without any loss on generality.

Lemma 2. For $\rho \in (0, 1)$

$$V_y(\rho) > V_{y'}(\rho) \quad \text{iff} \quad [Lik_y(\theta_1)/1, Lik_y(\theta_2)/0] \succ [Lik_{y'}(\theta_1)/1, Lik_{y'}(\theta_2)/0] \quad (16)$$

Theorem 1. Suppose ρ is a r.v. taking values in the unit interval. Then $V_y(\rho)$ stochastically dominates (first degree) $V_{y'}(\rho)$ iff

$$[Lik_y(\theta_1)/1, Lik_y(\theta_2)/0] \succ [Lik_{y'}(\theta_1)/1, Lik_{y'}(\theta_2)/0]$$

The order on canonical lotteries stipulated by axiom A5 is the order by first degree stochastic dominance of their expected payoffs if the prior is r.v. In Fig. 2, the lower curve is the graph for $V_{0.60}(\rho)$ (at $Y = .6$, the corresponding lottery is $[.3011/1, 1/0]$) and the upper curve is the graph for $V_{0.26}(\rho)$ ($[.5945/1, 1/0]$). This completes our justification for the five axioms.

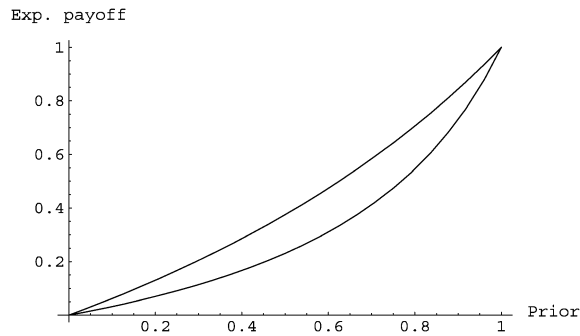


Fig. 2. Payoff functions for $y = .60$ (lower) and $y = .26$ (upper).

3.2. Binary utility

We will now proceed to study the preference relation satisfying the axioms. The following lemma shows that the set $\{\mathcal{L}/\sim\}$ of equivalent classes of lotteries wrt the indifference relation \sim is isomorphic to the set of canonical lotteries (\mathcal{L}_c).

Lemma 3. *If the preference relation \succeq on the set of lotteries \mathcal{L} satisfies axioms A1 though A5, then for each lottery there exists one and only one canonical lottery indifferent to it.*

The significance of Lemma 3 is that it reduces a comparison of lotteries to one of canonical lotteries that have a simple structure and a straightforward interpretation. We want to represent \succeq by a utility function so that a comparison of lotteries can be done through the calculation of their utilities. Our main idea here is to use as a utility scale a set that is isomorphic to the set of canonical lotteries. Let us define

$$\mathcal{U} \stackrel{\text{def}}{=} \{ \langle a, b \rangle \mid a, b \in [0, 1] \text{ and } \max(a, b) = 1 \} \quad (17)$$

In words, \mathcal{U} is the set of pair of numbers in the unit interval such that one of them is 1. A linear order \gg on \mathcal{U} (to distinguish from the order \geq on scalars) is defined as

$$\langle a, b \rangle \gg \langle a', b' \rangle \quad \text{iff} \quad \begin{cases} a = a' = 1 \text{ \& } b \leq b', \text{ or} \\ a = 1 \text{ \& } a' \leq 1, \text{ or} \\ a \geq a' \text{ \& } b = b' = 1 \end{cases} \quad (18)$$

Strict preference (\gg) and indifference ($=$) derivatives are also used. The special structure of \mathcal{U} allows a simplification of order definition given in Eq. (18). The proof of the following lemma is straightforward and is therefore omitted.

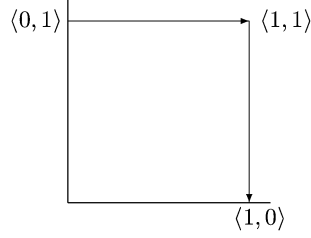
Lemma 4. *For $\langle a, b \rangle, \langle a', b' \rangle \in \mathcal{U}$*

$$\langle a, b \rangle \gg \langle a', b' \rangle \quad \text{iff} \quad (a > a') \vee (b < b') \quad (19)$$

$$\langle a, b \rangle \gg \langle a', b' \rangle \quad \text{iff} \quad (a \geq a') \wedge (b \leq b') \quad (20)$$

$$\langle a, b \rangle = \langle a', b' \rangle \quad \text{iff} \quad (a = a') \wedge (b = b') \quad (21)$$

We refer to \mathcal{U} equipped with order \gg as the *binary utility scale*. Roughly, we can interpret two components in a utility value as indices of goodness (first) and badness (second). One binary utility value is better than another if the goodness index of the former is higher than that of the latter or the badness index of the former is smaller than badness index of the latter. Note that this binary utility is a special case of the lexicographic utility [16]. Lemma 4 shows that two indices have symmetrical roles, no one has precedence over the other. One index is used as tie breaking in case equality holds for the other index.

Fig. 3. Binary utility scale \mathcal{U} .

To operate on binary utilities, we extend⁷ product and max operations to work with pairs as follows. For scalar α , β , π and γ

$$\alpha \cdot \langle \beta, \gamma \rangle \stackrel{\text{def}}{=} \langle \alpha \cdot \beta, \alpha \cdot \gamma \rangle \quad (22)$$

$$\max(\langle \alpha, \beta \rangle, \langle \gamma, \pi \rangle) \stackrel{\text{def}}{=} \langle \max(\alpha, \gamma), \max(\beta, \pi) \rangle \quad (23)$$

We have some properties of the extended *max* operation. The proof is straightforward and is therefore omitted.

Lemma 5. (i) \mathcal{U} is closed under max, i.e., $u, u' \in \mathcal{U}$ then $\max(u, u') \in \mathcal{U}$. (ii) max is monotone on each argument, for example, $\max(u, v) \geq \max(u, v')$ if $v \geq v'$.

3.3. Representation theorem

A utility function is a mapping from the set of lotteries into the utility scale $U : \mathcal{L} \rightarrow \mathcal{U}$. We say that a preference relation \succeq is *represented* by a utility function U whenever $L \succeq L'$ iff $U(L) \geq U(L')$. For a function f defined on set X , we let $f(X)$ denote $\{f(x) \mid x \in X\}$. We have the following theorem.

Theorem 2. \succeq on \mathcal{L} satisfies axioms A1 through A5 if and only if there exists a utility function $QU : \mathcal{L} \rightarrow \mathcal{U}$ representing \succeq such that $\langle 1, 0 \rangle, \langle 0, 1 \rangle \in QU(X)$ and

$$QU([\pi_1/L_1, \dots, \pi_k/L_k]) = \max_{1 \leq i \leq k} \{\pi_i \cdot QU(L_i)\} \quad (24)$$

While proposing axioms A1 through A5, we argue the rationale for each axiom separately, but not the consistency of the axiom system as a whole. The fact that the axiom system is represented by a well defined utility function proves that it is free from inconsistency. In particular, when L is a simple lottery, Eq. (24) can be rewritten as

$$QU([\pi_1/x_1, \dots, \pi_r/x_r]) = \max_{1 \leq i \leq r} \{\pi_i \cdot QU(x_i)\} \quad (25)$$

⁷ We have decided in favor of “overloading” operations *product* and *max* instead of creating new symbols. Hopefully, this slight abuse of notation does not lead to any confusion because the type of arguments will indicate which rule to apply.

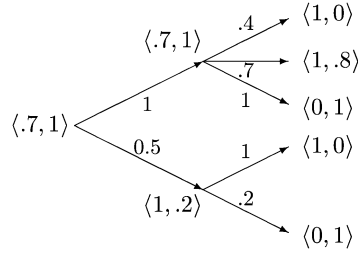


Fig. 4. Qualitative utility calculation.

The form of QU in (25) resembles the vNM expected utility expression. The expected utility of a probabilistic lottery p on X is defined as $U(p) \stackrel{\text{def}}{=} \sum_{1 \leq i \leq r} p(x_i) \cdot U(x_i)$.

Maximization in (25) plays the same role as its counterpart—addition—in the vNM expected utility. This similarity leads us to refer to QU also as *expected qualitative utility* function.

Fig. 4 illustrates QU calculation for a two-stage lottery. Assuming $QU(x_1) = \langle 1, 0 \rangle$, $QU(x_2) = \langle 1, .8 \rangle$ and $QU(x_3) = \langle 0, 1 \rangle$, a roll-back calculation shows that

$$QU([1/[.4/x_1, .7/x_2, 1/x_3], .5/[1/x_1, .2/x_3]]) = \langle .7, 1 \rangle$$

Although qualitative and vNM utilities share fundamental structure, it is important to emphasize that the qualitative utility is not just a simple translation of vNM utility in a new language. Not all properties of vNM utility hold for the qualitative version. Notably, the qualitative utility only satisfies weaker versions of independence and Archimedean properties [25].

Theorem 3. (a) Suppose $L_1, L_2, L_3 \in \mathcal{L}$, $\lambda, \mu \in [0, 1]$ and $\max(\lambda, \mu) = 1$. If $L_1 \succeq L_2$, then $[\lambda/L_1, \mu/L_3] \succeq [\lambda/L_2, \mu/L_3]$.

(b) Suppose $L_1, L_2, L_3 \in \mathcal{L}$ such that $L_1 \succ L_2 \succ L_3$. Then there exists $\lambda, \mu, \lambda', \mu' \in [0, 1]$, $\max(\lambda, \mu) = \max(\lambda', \mu') = 1$ and $\langle \lambda, \mu \rangle, \langle \lambda', \mu' \rangle \notin \{\langle 0, 1 \rangle, \langle 1, 0 \rangle\}$ such that $[\lambda/L_1, \mu/L_3] \succ L_2$ and $L_2 \succ [\lambda'/L_1, \mu'/L_3]$.

Note that property (a) does not hold for strict preference. That is, in general, we don't have $[\lambda/L_1, \mu/L_3] \succ [\lambda/L_2, \mu/L_3]$ if $L_1 \succ L_2$.

3.4. Related work

In the AI literature, a number of decision models that do not assume probability have been studied [7,24,37]. Brafman and Tennenholtz [7] characterize qualitative decision rules: *maximin*, *minimax regret* and *competitive ratios* and *maximax*. They show that these different decision criteria are equivalent in terms of representation power. These purely qualitative rules ignore the uncertainty relevant to choice problem. Smets [37] proposes a two-level decision model for Dempster–Shafer belief functions. At the *credal* level, an agent uses belief functions to represent and reason with uncertainty. When she needs to make decision she will translate a belief function into probability using *pignistic transformation*. Basically, this transformation allocates the probability mass that assigned to a

focus equally to each of its element. Smets' model is not able to handle ambiguity attitudes. For example, when applied for Ellsberg's paradox [15], this model produces an unintuitive solution. Halpern [24] studies a very general uncertainty measure called the *plausibility measure*. To make decisions, he defines an operation that maps the product of consequence domain and plausibility domain into a *valuation* domain. The order on valuation domain is defined by a decision rule. In [20], we also study a decision making model for Spohn's theory of epistemic beliefs. This uncertainty measure is closely related to extended likelihood, and can be interpreted as order-of-magnitude approximation of probabilities or as degrees of plain beliefs.

More relevant to this work is an approach to decision making with possibility theory proposed by Dubois et al. [10,13]. As noted earlier, a possibility function satisfies Eqs. (5), (6) that define an ELF. The main difference is about the conditioning operation. Dubois et al. use the ordinal conditioning defined by Eq. (8). The likelihood conditioning (or the numerical conditioning) is defined in Eq. (7). They distinguish two decision criteria: pessimistic and optimistic. For each decision criterion, there is an axiom system and a qualitative utility functional representing preference relation that satisfies the axioms. A detailed comparative analysis between our approach vs the approach argued by Dubois et al. is presented in [22]. We show that our approach, modified for ordinal conditioning, generalizes and unifies pessimistic and optimistic decision criteria.

In our framework, a decision maker's attitude toward ambiguity shows itself in her basic utility assignment for consequences in X . Recall that the indifference between a consequence and a (binary) utility value is determined through a likelihood gamble. We will see that her betting behavior encodes an interesting information, namely, her *attitude toward ambiguity*.

Suppose that our decision maker equates payoff x with canonical lottery $[\lambda/1, \mu/0]$. From the Bayesian decision theory point of view, this indifference means that the expected payoff $V_y(\rho)$ (with respect to prior ρ) of the likelihood gamble is equal to x . Substitute x for $V_y(\rho)$ and λ, μ for $Lik_y(\theta_1), Lik_y(\theta_2)$ into Eq. (13) and solve for ρ we find

$$\rho = \frac{x\mu}{(1-x)\lambda + x\mu} \quad (26)$$

From Eq. (26) we have

$$\text{logit}(\rho) = \text{logit}(x) - \ln(\lambda/\mu) \quad (27)$$

where $\text{logit}(z) \stackrel{\text{def}}{=} \ln(z/(1-z))$. We call ρ calculated by Eq. (27) an *implicit prior*⁸ for the obvious reason. When $\lambda = \mu = 1$ we have $\rho = x$. Therefore, ρ can be interpreted as the price the decision maker pays for a "fair" likelihood lottery $[1/1, 1/0]$.

Recall that in the context of a likelihood gamble, ρ is the probability, in Player's judgment, that the Arbiter selects θ_1 (on which \bar{x} is contracted) as the parameter value to generate the given observation. Thus inequality $\rho < 0.5$ can be interpreted as that the Player *a priori* deems the bad outcome is more likely than the good one. English language has a

⁸ It is necessary to note that the implicit prior value is unique for a single bet. A betting behavior that implies different implicit priors for different likelihood gambles can still be consistent with A5. For more details on the range of permissible priors, readers are referred to [19].

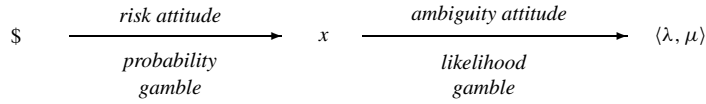


Fig. 5. Risk and ambiguity in transformation from money to binary utility.

specific name for the mental attitude that tends to emphasize adverse aspects—*pessimism*. Similarly, the opposite $\rho > 0.5$ can be argued as a manifestation of *optimism*. This definition of pessimism (optimism) is very different than the notion of pessimism (optimism)⁹ put forward by Dubois et al. [10] where it is a property of an axiom system attributed to a decision maker. In this paper, however, pessimism (optimism) is a property attributed to a basic utility assignment by a decision maker.

Another sensible terminology can be used for pessimism and optimism. A betting behavior is said to exhibit *ambiguity averse* ($\rho < 0.5$), *ambiguity neutral* ($\rho = 0.5$) or *ambiguity seeking* ($\rho > 0.5$) attitudes. It is useful to analyze similarity as well as distinction between the notion of ambiguity attitude and the established notion of risk attitude. The latter is a property of a utility function (money-to-utility conversion). Risk averse, risk neutral and risk seeking attitudes correspond to concavity, linearity and convexity of utility functions. In this paper, we do not explicitly consider the risk attitude issue because of the assumption that X measured in a utility currency. Fig. 5 illustrates the roles of risk and ambiguity attitudes in the transformation of monetary reward to a binary utility via unary utility.

In Economics and Statistics literature, a prominent axiomatic decision theory without additive probability has been studied by Schmeidler [33] and extended by Gilboa [23] and Sarin and Wakker [32]. The concept of co-monotonicity has a crucial role in Schmeidler's system. Two acts f and g are co-monotonic if for no two states s_1, s_2 $f(s_1) > f(s_2)$ and $g(s_2) > g(s_1)$. The independence property is required to hold only for set of co-monotonic acts. The preference relation that satisfies Schmeidler's axioms is represented by a *Choquet expected utility* function (CEU). A real-valued function $v: 2^\Omega \rightarrow [0, 1]$ is called *capacity* function if $v(\emptyset) = 0$, $v(\Omega) = 1$ and for $A \subseteq B \subseteq \Omega$ $v(A) \leq v(B)$. $CEU_v(f)$ is defined as follows. For simplicity, we assume that X is a set of reals, interpreted as utilities, and is ordered $x_0 > x_1 > \dots > x_m$. For an act f , $A_k^f \stackrel{\text{def}}{=} \{s \in \Omega \mid f(s) \geq x_k\}$ —set of states where f delivers x_k or better consequence. Obviously A_i^f are nested, i.e., $A_0^f \subseteq A_1^f \subseteq \dots \subseteq A_m^f = \Omega$

$$CEU_v(f) \stackrel{\text{def}}{=} x_0 v(A_0^f) + \sum_{i=1}^k x_i (v(A_i^f) - v(A_{i-1}^f)) \quad (28)$$

The main difference between Schmeidler's approach and our approach is in the role of a conditioning operation for uncertainty measure. CEU representation is the result of considering (i) separation of “utility” from “probability” and (ii) “functional representations

⁹ The definition of pessimism (optimism) by Dubois et al. requires that possibilistic lottery π is preferred to (less preferred than) lottery π' if the possibility of getting any consequence in π' is greater or equal to the possibility of getting that consequence in π , i.e., $\pi \succeq_{\text{pes}} \pi'$ ($\pi' \succeq_{\text{opt}} \pi$) if $\pi' \geq \pi$.

which are the sum of products of two numbers; one number has a “probability” interpretation and the other has a “utility” interpretation” [33, p. 584]. The updating operation is not explicitly considered for capacity function. In contrast, our approach starts with an uncertainty calculus with its properly defined updating rules and then develops a decision theory where utility is derived from “probability”. In this sense, although an ELF satisfies requirements for a capacity measure, it is not a special case of capacity because it has a well defined conditioning operation.

To see the difference between *CEU* and *QU*, let us calculate *CEU* for a canonical act with respect to ELF π . Suppose for simplicity $\bar{x} = 1$ and $\underline{x} = -1$. Since $A_0^a = A$ and $A_1^a = \Omega$, by Eq. (28)

$$CEU_{\pi}([A/\bar{x}, \bar{A}/\underline{x}]) = \pi(A) - (1 - \pi(A)) = 2\pi(A) - 1$$

First we observe that *CEU* for a canonical act does not depend on the capacity of event leading to the worst consequence. $v(\bar{A}) \neq 1 - v(A)$ because a capacity measure is non-additive. In contrast, A5 requires a comparison of likelihoods of getting worst consequence when the likelihoods of getting the best consequence are equal. This observation seems to suggest that *CEU* is not appropriate for existing (non-probabilistic) uncertainty calculi such as Dempster–Shafer belief functions [35], fuzzy possibility [41] and plausibility measures [24] that have well-defined updating rules.

The lack of an updating rule for capacity subjects *CEU* to the following criticism. Suppose v is a capacity measure, one can define its *dual* by $v'(A) \stackrel{\text{def}}{=} 1 - v(\bar{A}) \forall A \subseteq \Omega$. It is not difficult to see that v' is also a capacity measure. It is arguable that v and v' contain the same information because v is recoverable from v' . Despite visible symmetry between v and v' , rankings of acts by CEU_v and by $CEU_{v'}$ are different. In this sense, *CEU* is not sensitive to information. Such criticism is void for uncertainty calculi with well defined updating rules. The dual of a probability measure is itself. The dual of a possibility measure is a necessity measure but their updating rules are different. The same is also true for the dual pair of belief and plausibility functions in Dempster–Shafer theory. Thus, the measures are not symmetric despite being duals of each other.

4. Likelihood solution to statistical inference

4.1. Decision-theoretic approach to statistical inference

We will review the decision-theoretic approach to statistical inference. We assume as given the set of alternative actions denoted by \mathcal{A} , and the sample space of Y by \mathcal{Y} . A loss $V(a, \theta)$ measures the loss that arises if we take action a and the true value of the parameter is θ .¹⁰ A decision rule is a mapping $\delta: \mathcal{Y} \rightarrow \mathcal{A}$, that is for observation y the rule recom-

¹⁰ In terms of the decision problem definition (Section 3), any superset of $V(\mathcal{A}, \Omega)$, the set of possible loss values, could be the set of consequences X .

mends an action $\delta(y)$. The *risk function* of a decision rule δ at parameter value θ is defined as

$$R(\delta(Y), \theta) \stackrel{\text{def}}{=} E_{\theta} V(\delta(Y), \theta) = \int_{\mathcal{Y}} V(\delta(y), \theta) p_{\theta}(y) \quad (29)$$

The risk function measures the average loss by adopting the rule δ in case θ is the true value.

The further use of risk functions depends on how much information we assume is available. For each point in the parameter space, there is a value of the risk function. In case no a priori information about parameter exists, Wald [39] advocated the use of minimax rule which minimizes the worst risk that could be attained by a rule.

$$\delta_{\text{minimax}}^* = \arg \min_{\delta \in \Delta} \max_{\theta \in \Omega} R(\delta, \theta) \quad (30)$$

where Δ is the set of decision rules. δ^* is called the *minimax solution*.

If we assume, as the Bayesian school does, the existence of a prior distribution for the parameter, then the risk could be averaged out to one number called *Bayes risk*

$$r(\delta) = E_{\rho} R(\delta, \theta) = \int_{\Omega} R(\delta, \theta) \rho(\theta) \quad (31)$$

where ρ is prior distribution for θ . Then the optimal rule is one that minimizes the Bayes risk which is called the *Bayes solution*.

$$\delta_{\text{Bayes}, \rho}^* = \arg \min_{\delta \in \Delta} r(\delta) \quad (32)$$

Wald [39] pointed out there exists a prior distribution ρ^* called “the least favorable” for which the Bayes solution is the minimax solution. The term “Bayes” is justified by the fact that the solution is also obtained via a more intuitive route using Bayes theorem and the principle of minimizing expected loss. Given prior probability distribution ρ on Ω , for each data $y \in \mathcal{Y}$ a posterior probability on Ω is obtained via Bayes theorem

$$p(\theta | y) \propto p_{\theta}(y) \rho(\theta) \quad (33)$$

Denote by $a^p(y)$ the action that minimizes the expected loss given data y

$$a^p(y) = \arg \min_{a \in \mathcal{A}} \int_{\Omega} V(a, \theta) p(\theta | y) \quad (34)$$

Let us define a rule $\delta_p^*(y) \mapsto a^p(y)$, i.e., for each data y , rule δ_p^* delivers the action that minimizes the expected loss.

Lemma 6. δ_p^* is a Bayes solution, i.e., $r(\delta_p^*) = r(\delta_{\text{Bayes}, \rho}^*)$.

4.2. Likelihood solution

Without knowing prior ρ , we propose the following solution based on the logic that leads to δ_p^* . Each $y \in \mathcal{Y}$ is associated with an ELF Lik_y . An action together with an ELF

induce a likelihood lottery. Denote by $L_a(y)$ the lottery generated from action a and observation y . Likelihood lotteries are compared by QU . The optimal action given observation y is

$$a^{Lik}(y) = \arg \sup_{a \in \mathcal{A}} QU(L_a(y)) \quad (35)$$

where \sup^{11} is the operation taking maximum element according to the binary order \gg . We define a decision rule $\delta_{Lik}^*(y) \mapsto a^{Lik}(y)$ which assigns for each point in the sample space an action that maximizes the (qualitative) utility. We call such decision rule a *likelihood solution*.

There are two dimensions in which solutions $\delta_{minimax}^*$, δ_{Bayes}^* and δ_{Lik}^* can be compared. The first concerns information. One can differentiate two kinds of information relevant to the statistical inference problem: the extra-experiment information embodied by a prior probability and the experimental information embodied by a likelihood function. The Bayesian approach requires both kinds and utilizes them to arrive at the solution. In Wald's proposal, the risk function has no *special* role for the actually observed data and the prior probability is not considered. Thus both kinds of information are ignored. The likelihood solution does not assume knowing the prior, but it does make use of likelihood information provided by data in identifying the best action.

The concept of stochastic dominance provides another dimension for comparing the three solutions. Apart from FSD, stochastic dominance of second and higher degree are defined. For simplicity, following [6], assume r.v. are non-negative, i.e., cumulative distribution functions satisfy $F(0) = 0$. For cumulative distribution function F , define for any natural n

$$F_n(z) = \int_0^z F_{n-1}(x) dx \quad (36)$$

with notation $F_1 = F$. Suppose X, Y are two r.v. (we use the same symbols for their cumulative distribution function), we say X is preferred to Y according to n -degree stochastic dominance (write $XnSDY$) if $X_n(z) \leq Y_n(z)$ for $z \geq 0$. It is well known that (i) n -degree dominance implies all higher degree dominance and (ii) higher the degree, the greater relative importance is assigned to small value of r.v. Borch [6] shows that Wald's minimax rule is equivalent to stochastic dominance of infinite degree. The order satisfying vNM axioms can be viewed as "zero degree" stochastic dominance because it boils down to the comparison of numbers—expected utility values—that are, of course, singular r.v. Thus, we can arrange Bayes, likelihood and minimax solutions in an increasing order according to their SD degrees.

4.3. An illustrative example

The following example is adapted from [3]. The manufacturer of small travel clocks which are sold through a chain of department stores agrees to service any clock once only

¹¹ In contrast to *max* defined in Eq. (23) that operates on scalars \geq .

if it fails to operate satisfactorily in the first year of purchase. For any clock, a decision must be made on whether to merely clean it or replace the works, i.e., the set of actions $\mathcal{A} = \{a_1, a_2\}$ where a_1 denotes “clean the clock then replace the work if needed”, and a_2 denotes “immediately replace the works”.

Let us assume that there are only two possible faults, i.e., $\Omega = \{\theta_1, \theta_2\}$ where θ_1 means there is the need for cleaning and θ_2 means the clock has been physically damaged and the works need replacement. Utility and loss functions are given in the following table. The relationship between utility (u) and loss (V) is through equation $V = 1 - u$.

$u(a, \theta)$	θ_1	θ_2
a_1	.8	.3
a_2	.5	.5

$V(a, \theta)$	θ_1	θ_2
a_1	.2	.7
a_2	.5	.5

The loss table is roughly estimated from the fact that cleaning a clock costs \$0.20 and replacing the works costs \$0.50. If the policy is to replace the works for every clock needing service then the cost is \$0.50 no matter which problem is present. If the policy is to clean a clock first, if the state is θ_1 then the service costs \$0.20, however in the case of physical damage then cleaning alone obviously does not fix the problem and the manufacturer ends up replacing the works also. Thus the total cost is \$0.70.

The manufacturer can ask the customer to provide a symptom of malfunction when a clock is sent to the service center. Assume the sample space $\mathcal{Y} = \{y_1, y_2, y_3\}$ where y_1 means “the clock has stopped operating”, y_2 —“the clock is erratic in timekeeping” and y_3 —“clock can only run for a limited period”. Such information gives some indication about θ that is expressed in terms of likelihood

$lik_y(\theta)$ or $p_\theta(y)$	y_1	y_2	y_3
θ_1	.1	.4	.5
θ_2	.7	.2	.1

For each point in the sample space, you can either choose a_1 or a_2 , so there are 8 possible decision rules in total. Each decision rule specifies an action given an observation.

$\delta_j(y_i)$	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
y_1	a_1	a_1	a_1	a_1	a_2	a_2	a_2	a_2
y_2	a_1	a_1	a_2	a_2	a_1	a_1	a_2	a_2
y_3	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2

We calculate the risk function values for each rule and parameter value in the following table

R_{ij}	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
θ_1	.2	.35	.32	.47	.23	.38	.35	.50
θ_2	.7	.68	.66	.64	.56	.54	.52	.50

Notice that there is no rule which is superior to all other for both values of θ . Wald’s minimax solution is δ_8 . If we assume prior distribution of θ then we can calculate the Bayes risks for the rules. For example if prior probability $p(\theta_1) = .7$, then Bayes risks $r_{.7}(\delta_i)$ are

δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
.35	.449	.442	.541	.329	.428	.401	.50

Thus, δ_5 is the Bayes solution. The following sensitivity analysis shows how Bayes solutions depend on prior $p(\theta_1)$

Bayes solution	When
δ_8	$p(\theta_1) \leq .118$
δ_7	$.118 \leq p(\theta_1) \leq .250$
δ_5	$.250 \leq p(\theta_1) \leq .824$
δ_1	$.824 \leq p(\theta_1)$

In our approach, suppose unary utility is translated into binary utility according to the following table. The table is obtained assuming ambiguity neutrality. For example, to find a binary utility equivalent to 0.8, plugging $x = 0.8$ and implicit prior $\rho = 0.5$ into Eq. (27) we have $\text{logit}(0.8) = \ln(\lambda/\mu) + \text{logit}(0.5)$. Therefore, $\lambda/\mu = 4$. Since $\max(\lambda, \mu) = 1$ we have $\lambda = 1$ and $\mu = .25$. Thus $0.8 \sim \langle 1, 0.25 \rangle$.

Unary utility	Binary utility
.8	$\langle 1, .25 \rangle$
.5	$\langle 1, 1 \rangle$
.3	$\langle .43, 1 \rangle$

Given y_1 , the ELF is $Lik_{y_1}(\theta_1) = .14$ and $Lik_{y_1}(\theta_2) = 1$. Action a_1 corresponds to lottery $L_{a_1}(y_1) = [.14/\langle 1, .25 \rangle, 1/\langle .43, 1 \rangle]$

$$\begin{aligned} QU(L_{a_1}(y_1)) &= \max\{.14\langle 1, .25 \rangle, 1\langle .43, 1 \rangle\} \\ &= \max\{\langle .14, .035 \rangle, \langle .43, 1 \rangle\} = \langle .43, 1 \rangle \end{aligned}$$

Action a_2 is associated with lottery $L_{a_2}(y_1) = [.14/\langle 1, 1 \rangle, 1/\langle 1, 1 \rangle]$ and $QU(L_{a_2}(y_1)) = \langle 1, 1 \rangle$. Thus, $a_2 \succ_{y_1} a_1$. Given y_2 , $Lik_{y_2}(\theta_1) = 1$ and $Lik_{y_2}(\theta_2) = .5$. So, $QU(L_{a_1}(y_2)) = \langle 1, .5 \rangle$ and $QU(L_{a_2}(y_2)) = \langle 1, 1 \rangle$. Thus, $a_1 \succ_{y_2} a_2$. Given y_3 , $Lik_{y_3}(\theta_1) = 1$ and $Lik_{y_3}(\theta_2) = .2$. $QU(L_{a_1}(y_3)) = \langle 1, .25 \rangle$ and $QU(L_{a_2}(y_3)) = \langle 1, 1 \rangle$. Thus, $a_1 \succ_{y_3} a_2$. In summary, the likelihood solution is δ_5 .

It is also interesting to know how likelihood solutions depend on the ambiguity attitude. It can be shown

Likelihood solution	Implicit prior
δ_7	$\rho \leq 0.3333$
δ_5	$0.3334 \leq \rho$

Let us make few informal comments. Depending on ambiguity attitude, the likelihood solution can be δ_7 or δ_5 while the minimax solution is δ_8 . As we noted, the latter ignores the uncertainty generated by an observation while the former does not.

If the prior probability is a value in interval $[0.250, 0.824]$, then the Bayes solution is δ_5 the same as the likelihood solution when the ambiguity attitude ρ is in $[0.3334, 1]$. If

prior probability is known, one can argue that Bayes solution is *the* optimal one. However, the “optimality” of the Bayes solution does not come without cost. The prior probability requirement can be satisfied either at some monetary cost (doing research, or buying from those who have) that should have been deducted from the benefit of having the optimal action. Alternatively, the decision maker can just assume some prior distribution in an *ad hoc* manner. This, however, would compromise the claimed optimality. One can extend the concept of Bayes solution by including a sensitivity analysis. This certainly helps decision maker by providing a frame of reference. But sensitivity analysis itself does not constitute any basis for knowing the prior probability. Moreover, when the parameter space is large, a multiple way sensitivity analysis may be very difficult.

The behavior comparison of the likelihood solution and the Bayes solution when ρ and $p(\theta_1)$ vary is more interesting. When $p(\theta_1)$ increases from 0 to 1, the Bayes solution moves from δ_8 to δ_7 to δ_5 and to δ_1 . When ρ increases from 0 to 1, the likelihood solution moves from δ_7 to δ_5 . Thus, the likelihood solutions do not include δ_8, δ_1 that correspond to Bayes solutions at extreme values of prior $p(\theta_1)$. In particular, δ_7 chooses a_1 and δ_8 chooses a_2 at y_3 . δ_1 chooses a_1 and δ_5 chooses a_2 at y_1 . There are two factors that ameliorate this discrepancy. First, extreme values of $p(\theta_1)$ are far from non-informative priors which can be expected in a situation of no prior information. In other words, by selecting $p(\theta_1)$ close to 0 or 1, the decision maker has some (or is willing to assume) significant a priori information that would make the comparison with likelihood solution void. Second, at extreme values of implicit prior ρ , the utilities of a_1 and a_2 (at any observation) are very close. For example, if $\rho = 0.1$ then $QU(L_{a_1}(y_3)) = \langle 1, 0.052 \rangle$ and $QU(L_{a_2}(y_3)) = \langle 1, 0.111 \rangle$. At non-extreme cases the likelihood solution and the Bayes solution basically agree. Decision rules δ_7, δ_5 are selected in both approaches for large segments of values $p(\theta_1)$ and ρ .

This pattern is comforting but one should be careful not to draw too much from it. Specifically, one should not use this agreement as a justification for the likelihood solution. As we pointed out, axioms A1 to A5 on which likelihood solution is based, are structurally similar to those used by Luce and Raiffa [29] to justify the expected utility maximization principle which ultimately is the basis for Bayes solutions. Thus, at a foundational level, optimality of likelihood solution could be justified in the same way as the optimality for Bayes solution although the two optimality concepts are obviously different. It can be argued that the question of which optimality is more appropriate depends on how much information is available.

Although $p(\theta_1)$ and ρ have similar roles on the behavior of solutions, they have very different semantics. $p(\theta_1)$ is a specification of prior probability on the parameter space. Therefore, it gauges the a priori information (empirical or assumed) about parameters. In practice, when this space has many elements, the specification would be much more difficult and sensitivity analysis becomes more complex. ρ , on the other hand, gauges the attitude toward ambiguity of a decision maker. It is determined, independently to the parameter space, on simple situations of two hypotheses and fixed rewards.

5. Summary and conclusion

In this paper, we develop a decision theory that directly utilizes likelihood information without assuming the knowledge of prior probability. We extend likelihood functions as the uncertainty measure pertaining to a statistical inference problem. This extension, conforming to the maximum likelihood methods, defines the likelihood for a set of parameter values to be the maximum likelihood over elements of the set.

Our approach is axiomatic. The axioms considered are similar in spirit to those of the linear utility theory, but strictly different in several important aspects. We describe a betting behavior based on likelihood rather than on probability. This behavior satisfies the stochastic dominance principle. We prove a representation theorem for preference relation over likelihood lotteries using a newly developed concept of binary utility.

Applied to the statistical inference problem, our theory suggests a new solution that picks an action by maximizing expected qualitative utility. This solution is sandwiched between Wald's minimax solution and the Bayes solution in terms of information use and demand. It makes use of uncertainty information that is ignored by the minimax solution but does not require a prior probability as the Bayes solution does.

Appendix A. Proofs of lemmas and theorems

Proof of Lemma 2. By Eq. (15), $V_y(\rho) > V_{y'}(\rho)$ iff $P_\rho(\theta = \theta_1 | y) > P_\rho(\theta = \theta_1 | y')$. By Eq. (13), $P_\rho(\theta = \theta_1 | y) > P_\rho(\theta = \theta_1 | y')$ iff

$$\frac{Lik_y(\theta_1)}{Lik_y(\theta_2)} > \frac{Lik_{y'}(\theta_1)}{Lik_{y'}(\theta_2)} \quad (\text{A.1})$$

Because $\max(Lik_y(\theta_1), Lik_y(\theta_2)) = \max(Lik_{y'}(\theta_1), Lik_{y'}(\theta_2)) = 1$, there are 4 cases to consider. Eq. (A.1) excludes the case where $Lik_y(\theta_2) = Lik_{y'}(\theta_1) = 1$. For 3 remaining cases, we have (a) If $Lik_y(\theta_1) = Lik_{y'}(\theta_1) = 1$, Eq. (A.1) holds iff $Lik_y(\theta_2) < Lik_{y'}(\theta_2)$; (b) If $Lik_y(\theta_2) = Lik_{y'}(\theta_2) = 1$, Eq. (A.1) holds iff $Lik_y(\theta_1) > Lik_{y'}(\theta_1)$; (c) If $Lik_y(\theta_1) = Lik_{y'}(\theta_2) = 1$, Eq. (A.1) holds iff either $Lik_y(\theta_2) < 1$ or $Lik_{y'}(\theta_1) < 1$. By Eq. (9), we have Eq. (A.1) holds iff $[Lik_y(\theta_1)/1, Lik_y(\theta_2)/0] \succ [Lik_{y'}(\theta_1)/1, Lik_{y'}(\theta_2)/0]$. \square

Proof of Theorem 1. (\Rightarrow) For any $v \in (0, 1)$, let us denote the roots of equations $V_y(\rho) = v$ and $V_{y'}(\rho) = v$ by ρ_v and ρ'_v respectively, i.e., $V_y(\rho_v) = v$ and $V_{y'}(\rho'_v) = v$. If $[Lik_y(\theta_1)/1, Lik_y(\theta_2)/0] \succ [Lik_{y'}(\theta_1)/1, Lik_{y'}(\theta_2)/0]$ then by Eq. (16) $V_y(\rho_v) > V_{y'}(\rho_v)$. Therefore, $V_{y'}(\rho'_v) > V_{y'}(\rho_v)$. Because $V_{y'}(\rho)$ is strictly increasing, we infer $\rho_v < \rho'_v$. Since $V_y(\rho)$ and $V_{y'}(\rho)$ are increasing, $P(V_y(\rho) \leq v) = P(\rho \leq \rho_v)$ and $P(V_{y'}(\rho) \leq v) = P(\rho \leq \rho'_v)$. Because $\rho_v < \rho'_v$, $P(V_y(\rho) \leq v) \leq P(V_{y'}(\rho) \leq v)$. This last inequality means $V_y(\rho) D_1 V_{y'}(\rho)$.

(\Leftarrow) If for all $0 < x < 1$, $V_y(x) \leq V_{y'}(x)$, then assumption $V_y(\rho) D_1 V_{y'}(\rho)$ is violated. Otherwise, Eq. (16) implies that

$$[Lik_y(\theta_1)/1, Lik_y(\theta_2)/0] \succ [Lik_{y'}(\theta_1)/1, Lik_{y'}(\theta_2)/0] \quad \square$$

Proof of Lemma 3. We prove the existence of indifferent canonical lottery by induction on the depth of lottery trees. For a constant lottery (of depth 0), because of A4, each consequence x_i is indifferent to a canonical lottery s_i . Let assume $x_i \sim s_i = [\kappa_{i1}/\bar{x}, \kappa_{ir}/\underline{x}]$ for $1 \leq i \leq r$.

A lottery of depth 1 is a simple lottery. If it is a canonical lottery, by reflexivity, a canonical lottery is indifferent to itself. For a simple lottery $L = [\pi_1/x_1, \pi_2/x_2, \dots, \pi_r/x_r]$, by A3, $L \sim L_1$ where $L_1 = [\pi_1/s_1, \pi_2/s_2, \dots, \pi_r/s_r]$. L_1 can be reduced to a canonical lottery L_2 such that $L_1 \sim L_2$ as follows. Let us write a canonical lottery s_i in the form $[\kappa_{i1}/\bar{x}, \kappa_{i2}/x_2, \dots, \kappa_{ir}/\underline{x}]$ with $\kappa_{ij} = 0$ for $2 \leq j \leq r-1$. By A2, $L_1 \sim L_2$ where $L_2 = [\kappa_1/\bar{x}, \kappa_2/x_2, \dots, \kappa_r/\underline{x}]$ with $\kappa_j = \max\{\pi_i \cdot \kappa_{ij} \mid 1 \leq i \leq k\}$. Since $\kappa_{ij} = 0$ for $2 \leq j \leq r-1$, we have $\kappa_j = 0$ for $2 \leq j \leq r-1$. Thus, L_2 is a canonical lottery. By transitivity, $L \sim L_2$.

Suppose for any lottery of depth not greater than n , there is a canonical lottery indifferent to it. For a lottery L of depth $n+1$. This lottery is a compound lottery whose consequences are lotteries of depth not greater than n . Because of induction hypothesis, each consequence of L is indifferent to a canonical lottery. By substitutability, L is indifferent to a compound lottery of depth 2 which, in turn, is indifferent to a canonical lottery by induction hypothesis. By transitivity, L is indifferent to some canonical lottery.

Finally, we have to show that there is only one canonical lottery indifferent to a given lottery. Suppose there are two canonical lotteries $s_1, s_2 \in \mathcal{L}_c$ such that $s_1 \sim L$ and $s_2 \sim L$. By A1, we have $s_1 \sim s_2$. But by A5, this is possible only if $s_1 = s_2$. \square

Proof of Theorem 2. (\Rightarrow) Suppose \succeq satisfies the axioms. We will show that there exists a function $QU: \mathcal{L} \rightarrow \mathcal{U}$ that satisfies Eq. (24) and represents \succeq . We construct function QU as follows. For a canonical lottery, define $QU([\lambda/\bar{x}, \mu/\underline{x}]) = \langle \lambda, \mu \rangle$. Obviously, $\langle 1, 0 \rangle, \langle 0, 1 \rangle \in QU(X)$. For any lottery L , by Lemma 3, there exists a unique canonical s such that $L \sim s$, we set $QU(L) = QU(s)$. Obviously, QU is well defined. By Eqs. (9) and (18), for canonical lotteries s, s' we have $s \succeq s'$ iff $QU(s) \succeq QU(s')$. That fact together with Lemma 3 and the way by which QU is defined allow us to conclude QU represents \succeq .

Now, we will show that QU satisfies Eq. (24). Consider depth-one lottery $L = [\pi_1/x_1, \dots, \pi_r/x_r]$. By A4, each consequence x_i is indifferent to a canonical lottery, say $s_i = [\kappa_{i1}/\bar{x}, \kappa_{ir}/\underline{x}]$. Therefore, $QU(x_i) = \langle \kappa_{i1}, \kappa_{ir} \rangle$. Consider lottery $L' = [\pi_1/s_1, \pi_2/s_2, \dots, \pi_r/s_r]$. From A3, $L \sim L'$. Using A2 and the argument in the proof of Lemma 3, we have L' is indifferent to canonical lottery $s = [\kappa_1/\bar{x}, \kappa_r/\underline{x}]$ where

$$\kappa_l = \max_{1 \leq i \leq r} \{\pi_i \cdot \kappa_{il}\} \quad \text{where } l \in \{1, r\} \quad (\text{A.2})$$

Therefore, on one hand $QU(L) = QU(L') = QU(s) = \langle \kappa_1, \kappa_r \rangle$. On the other hand,

$$\begin{aligned} \max_{1 \leq i \leq r} \{\pi_i \cdot QU(x_i)\} &= \max_{1 \leq i \leq r} \{\pi_i \cdot \langle \kappa_{i1}, \kappa_{ir} \rangle\} = \max_{1 \leq i \leq r} \{\langle \pi_i \cdot \kappa_{i1}, \pi_i \cdot \kappa_{ir} \rangle\} \\ &= \langle \max_{1 \leq i \leq r} \{\pi_i \cdot \kappa_{i1}\}, \max_{1 \leq i \leq r} \{\pi_i \cdot \kappa_{ir}\} \rangle = \langle \kappa_1, \kappa_r \rangle \end{aligned}$$

Thus, we show $QU(L) = \max_{1 \leq i \leq r} \{\pi_i \cdot QU(x_i)\}$. By induction on lottery's depth, we can prove this property for any lottery.

(\Leftarrow) Suppose \succeq_q is represented by QU that satisfies Eq. (24) and $\langle 1, 0 \rangle, \langle 1, 0 \rangle \in QU(X)$, i.e., $L \succeq_q L'$ iff $QU(L) \gg QU(L')$. We show that \succeq_q satisfies axioms A1 through A5. A1 is satisfied because relation \gg defined on \mathcal{U} by Eq. (18) is reflexive, complete and transitive.

Let $L = [\pi_1/L_1, \dots, \pi_i/L_i, \dots, \pi_k/L_k]$ and $L' = [\pi_1/L_1, \dots, \pi_i/L'_i, \dots, \pi_k/L_k]$. Assume $L_i \sim_q L'_i$. By definition of \succeq_q , it means $QU(L_i) = QU(L'_i)$. Apply Eq. (24) twice for compound lotteries L, L' . We see that the right-hand sides are identical. So the left-hand sides which are $QU(L)$ and $QU(L')$ must be equal. By definition of \succeq_q , we have $L \sim_q L'$. Thus, A3 is satisfied.

Let $L = [\pi_1/L_1, \pi_2/L_2, \dots, \pi_k/L_k]$ where $L_i = [\kappa_{i1}/x_1, \kappa_{i2}/x_2, \dots, \kappa_{ir}/x_r]$ for $1 \leq i \leq k$. Let us assume $QU(x_j) = \langle \lambda_j, \mu_j \rangle$ for $1 \leq j \leq r$. Apply Eq. (24) for L_i and then L ,

$$\begin{aligned} QU(L_i) &= \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \langle \lambda_j, \mu_j \rangle \} = \max_{1 \leq j \leq r} \{ \langle \kappa_{ij} \cdot \lambda_j, \kappa_{ij} \cdot \mu_j \rangle \} \\ &= \langle \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \lambda_j \}, \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \mu_j \} \rangle \\ QU(L) &= \max_{1 \leq i \leq k} \{ \pi_i \cdot \langle \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \lambda_j \}, \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \mu_j \} \rangle \} \\ &= \max_{1 \leq i \leq k} \{ \langle \pi_i \cdot \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \lambda_j \}, \pi_i \cdot \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \mu_j \} \rangle \} \\ &= \langle \max_{1 \leq i \leq k} \{ \pi_i \cdot \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \lambda_j \} \}, \max_{1 \leq i \leq k} \{ \pi_i \cdot \max_{1 \leq j \leq r} \{ \kappa_{ij} \cdot \mu_j \} \} \rangle \\ &= \langle \max_{1 \leq i \leq k} \{ \max_{1 \leq j \leq r} \{ \pi_i \cdot \kappa_{ij} \cdot \lambda_j \} \}, \max_{1 \leq i \leq k} \{ \max_{1 \leq j \leq r} \{ \pi_i \cdot \kappa_{ij} \cdot \mu_j \} \} \rangle \\ &= \langle \max_{1 \leq j \leq r} \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \cdot \lambda_j \}, \max_{1 \leq j \leq r} \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \cdot \mu_j \} \rangle \end{aligned}$$

Let us consider the simple lottery mentioned in A2: $L_s = [\kappa_1/x_1, \dots, \kappa_r/x_r]$ where

$$\kappa_j = \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \}$$

Apply Eq. (24) for L_s , we have

$$\begin{aligned} QU(L_s) &= \max_{1 \leq j \leq r} \{ \kappa_j \cdot \langle \lambda_j, \mu_j \rangle \} = \max_{1 \leq j \leq r} \{ \langle \kappa_j \cdot \lambda_j, \kappa_j \cdot \mu_j \rangle \} \\ &= \max_{1 \leq j \leq r} \{ \langle \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \cdot \lambda_j \}, \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \cdot \mu_j \} \rangle \} \\ &= \langle \max_{1 \leq j \leq r} \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \cdot \lambda_j \}, \max_{1 \leq j \leq r} \max_{1 \leq i \leq k} \{ \pi_i \cdot \kappa_{ij} \cdot \mu_j \} \rangle \end{aligned}$$

Comparing the last expressions, we have $QU(L) = QU(L_s)$. By definition of \succeq_q , $L \sim_q L_s$. Thus, A2 is satisfied.

Denote by \bar{x}, \underline{x} the elements of X such that $QU(\bar{x}) = \langle 1, 0 \rangle$ and $QU(\underline{x}) = \langle 0, 1 \rangle$. By Eq. (18) and definition of \succeq_q , we have $\bar{x} \succeq_q x$ and $x \succeq_q \underline{x}$ for all $x \in X$. For any canonical lottery $[\lambda/\bar{x}, \mu/\underline{x}]$ where $\max\{\lambda, \mu\} = 1$, by Eq. (24) we have $QU([\lambda/\bar{x}, \mu/\underline{x}]) = \max\{\lambda \cdot \langle 1, 0 \rangle, \mu \cdot \langle 0, 1 \rangle\} = \max\{\langle \lambda, 0 \rangle, \langle 0, \mu \rangle\} = \langle \lambda, \mu \rangle$. Thus, by Eq. (18) we conclude that A5 is satisfied.

Finally, for $x \in X$, let assume $QU(x) = \langle \lambda, \mu \rangle$. By above argument, we have $QU(x) = QU([\lambda/\bar{x}, \mu/\underline{x}])$. By definition of \succeq_q we infer $x \sim_q [\lambda/\bar{x}, \mu/\underline{x}]$. Thus, A4 is satisfied. \square

Proof of Theorem 3. Let us assume $QU(L_i) = \langle \lambda_i, \mu_i \rangle$ for $i = 1, 2, 3$.

(a) If $L_1 \sim L_2$, then the conclusion is a result of substitutability. Suppose $L_1 \succ L_2$. This means $\lambda_1 \geq \lambda_2$ & $\mu_1 \leq \mu_2$ and at least one of them is a strict relation. Applying Theorem 2, we have

$$\begin{aligned} QU([\lambda/L_1, \mu/L_3]) &= \langle \max(\lambda.\lambda_1, \mu.\lambda_3), \max(\lambda.\mu_1, \mu.\mu_3) \rangle \\ QU([\lambda/L_2, \mu/L_3]) &= \langle \max(\lambda.\lambda_2, \mu.\lambda_3), \max(\lambda.\mu_2, \mu.\mu_3) \rangle \end{aligned}$$

So, $\max(\lambda.\lambda_1, \mu.\lambda_3) \geq \max(\lambda.\lambda_2, \mu.\lambda_3)$, $\max(\lambda.\mu_1, \mu.\mu_3) \leq \max(\lambda.\mu_2, \mu.\mu_3)$. This means $QU([\lambda/L_1, \mu/L_3]) \geq QU([\lambda/L_2, \mu/L_3])$. By representation theorem

$$[\lambda/L_1, \mu/L_3] \succeq [\lambda/L_2, \mu/L_3].$$

(b) $L_1 \succ L_2 \succ L_3$ means $\lambda_1 \geq \lambda_2 \geq \lambda_3$ & $\mu_1 \leq \mu_2 \leq \mu_3$ and for indices $1 \leq i < j \leq 3$ either $\lambda_i > \lambda_j$ or $\mu_i < \mu_j$. We will identify $\lambda, \mu \geq 0$ satisfying $\max(\lambda, \mu) = 1$ such that $[\lambda/L_1, \mu/L_3] \succ L_2$. Since $QU([\lambda/L_1, \mu/L_3]) = \langle \max(\lambda.\lambda_1, \mu.\lambda_3), \max(\lambda.\mu_1, \mu.\mu_3) \rangle$, the requirement will be satisfied if either $\max(\lambda.\lambda_1, \mu.\lambda_3) > \lambda_2$ or $\max(\lambda.\mu_1, \mu.\mu_3) < \mu_2$. If $\lambda_1 > \lambda_2$, choosing $\lambda = 1$ will satisfy the former inequality. Otherwise $\lambda_1 = \lambda_2$, we have then $\mu_1 < \mu_2$. We choose $\lambda = 1$ and μ strictly positive and small enough so that $\mu.\mu_3 < \mu_2$. Thus $\max(\lambda.\mu_1, \mu.\mu_3) < \mu_2$. Similarly, we can choose λ', μ' so that $L_2 \succ [\lambda'/L_1, \mu'/L_3]$. \square

Proof of Lemma 6. We need to show that the Bayes risk for δ_p^* is minimal, i.e.,

$$\forall \delta \in \Delta \quad r(\delta_p^*) \leq r(\delta) \quad (\text{A.3})$$

Substitute Eq. (29) into Eq. (31), we have

$$r(\delta) = \int_{\Omega} \int_{\mathcal{Y}} V(\delta(y), \theta) p_{\theta}(y) \rho(\theta) = \int_{\mathcal{Y}} \int_{\Omega} V(\delta(y), \theta) p_{\theta}(y) \rho(\theta) \quad (\text{A.4})$$

It would be enough to show that $\forall y \in \mathcal{Y}$,

$$\int_{\Omega} V(\delta_p^*(y), \theta) p_{\theta}(y) \rho(\theta) = \int_{\Omega} V(a^p(y), \theta) p_{\theta}(y) \rho(\theta) \leq \int_{\Omega} V(\delta(y), \theta) p_{\theta}(y) \rho(\theta)$$

The last inequality is an implication of Eqs. (33) and (34). \square

References

- [1] Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrox, F. Caski (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973, p. 267.
- [2] M. Allais, The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School, in: M. Allais, O. Hagen (Eds.), Expected Utility Hypotheses and the Allais Paradox, Reidel, Dordrecht, 1979, pp. 27–145. The English translation of “Fondements d’une Theorie Positive des Choix Comportant un Risque et Critique des Postulats et Axiomes de L’Ecole Americaine”, *Econometrie, Colloques Internationaux du Centre National de la Recherche Scientifique*, Paris, vol. XL, 1953, pp. 275–332.
- [3] V. Barnett, Comparative Statistical Inference, third edition, Wiley, New York, 1999.

- [4] D. Basu, I. Ghosh, *Statistical Information and Likelihood: A Collection of Critical Essays by Dr. Basu*, Lecture Notes in Statistics, Springer, New York, 1988.
- [5] J.O. Berger, R.L. Wolpert, *The Likelihood Principle*, second edition, Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, CA, 1988.
- [6] K. Borch, Utility and stochastic dominance, in: M. Allais, O. Hagen (Eds.), *Expected utility hypotheses and the Allais paradox*, Reidel, Dordrecht, 1979, pp. 193–202.
- [7] R.I. Brafman, M. Tennenholtz, An axiomatic treatment of three qualitative decision criteria, *J. ACM* 47 (3) (2000) 452–483.
- [8] D.R. Cox, D.V. Hinkley, *Theoretical Statistics*, Chapman and Hall, London, 1974.
- [9] A. Dempster, A generalization of Bayesian inference, *J. Roy. Statist. Soc. Ser. B* 30 (1968) 205–247, with discussion.
- [10] D. Dubois, L. Godo, H. Prade, A. Zapico, On the possibilistic decision model: from decision under uncertainty to case-based decision, *Internat. J. Uncertainty Fuzziness Knowledge-based Syst.* 7 (6) (1999) 631–670.
- [11] D. Dubois, S. Moral, H. Prade, A semantics for possibility theory based on likelihoods, *J. Math. Anal. Appl.* 205 (1997) 359–380.
- [12] D. Dubois, T.H. Nguyen, H. Prade, Possibility theory, probability and fuzzy sets, in: D. Dubois, H. Prade (Eds.), *Handbook of Fuzzy Sets Series*, Kluwer Academic, Boston, 2000, pp. 344–438.
- [13] D. Dubois, H. Prade, Possibility theory as basis for qualitative decision theory, in: *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Quebec, vol. 2, Morgan Kaufmann, San Mateo, CA, 1995, pp. 1924–1930, <ftp://ftp.irit.fr/pub/IRIT/RPDMP/PTBQDT.ps.gz>.
- [14] A.W.F. Edwards, *Likelihood*, Cambridge University Press, Cambridge, 1972.
- [15] D. Ellsberg, Risk, ambiguity and the Savage's axioms, *Quart. J. Econ.* 75 (4) (1961) 643–669.
- [16] P.C. Fishburn, Lexicographic orders, utilities and decision rules: A survey, *Management Sci.* 20 (11) (1974) 1442–1471.
- [17] P.C. Fishburn, *Nonlinear Preference and Utility Theory*, The Johns Hopkins University Press, Baltimore, 1988.
- [18] R.A. Fisher, On the mathematical foundation of theoretical statistics, *Philos. Trans. Royal Soc. London A* 222 (1922) 309–368. Reprinted in: J.H. Bennett (Ed.), *Collected Papers of R.A. Fisher*, vol. 1, University of Adelaide, 1971.
- [19] P.H. Giang, *A decision theory for non-probabilistic uncertainty and its applications*, PhD thesis, University of Kansas, Lawrence, Kansas, 2003.
- [20] P.H. Giang, P.P. Shenoy, A qualitative linear utility theory for Spohn's theory of epistemic beliefs, in: C. Boutilier, M. Goldszmidt (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, San Francisco, CA, Morgan Kaufmann, San Mateo, CA, 2000, pp. 220–229.
- [21] P.H. Giang, P.P. Shenoy, Statistical decisions using likelihood information without prior probabilities, in: A. Darwiche, N. Friedman (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, San Francisco, CA, Morgan Kaufmann, San Mateo, CA, 2002, pp. 170–178.
- [22] P.H. Giang, P.P. Shenoy, Two axiomatic approaches to decision making using possibility theory, *European J. Oper. Res.* 162 (2) (2005) 450–467.
- [23] I. Gilboa, Expected utility with purely subjective non-additive probabilities, *J. Math. Econ.* 16 (1987) 65–88.
- [24] J.Y. Halpern, *Reasoning about Uncertainty*, MIT Press, Cambridge, MA, 2003.
- [25] N.E. Jensen, An introduction to Bernoullian utility theory I: Utility functions, *Swedish J. Econ.* 69 (1967) 163–183.
- [26] W.C.M. Kallenberg, Asymptotic Optimality of Likelihood Ratio Tests in Exponential Families, *Mathematical Centre Tracts*, vol. 77, Mathematisch Centrum, Amsterdam, 1978.
- [27] E.L. Lehmann, G. Casella, *Theory of Point Estimation*, second edition, Springer Texts in Statistics, Springer, New York, 1998.
- [28] H. Levy, Stochastic dominance and expected utility: Survey and analysis, *Management Sci.* 38 (4) (1992) 555–593.
- [29] R.D. Luce, H. Raiffa, *Games and Decision*, Wiley, New York, 1957.
- [30] J. Neyman, E.S. Pearson, On the use and interpretation of certain test criteria for purposes of statistical inference, part 1, *Biometrika* 20A (1928) 175–240. In: J. Neyman, E.S. Pearson, *Joint Statistical Papers*, University of California Press, 1967.

- [31] W.B. Poland, R.D. Shachter, Three approaches to probability model selection, in: R. Lopez de Mantaras, D. Poole (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference (UAI-94)*, Morgan Kaufmann, San Mateo, CA, 1994.
- [32] R. Sarin, P. Wakker, A simple axiomatization of nonadditive expected utility, *Econometrica* 60 (6) (1992) 1255–1272.
- [33] D. Schmeidler, Subjective probability and expected utility without additivity, *Econometrica* 57 (3) (1989) 571–587.
- [34] G. Schwarz, Estimating the dimension of a model, *Ann. Statist.* 6 (1978) 461–464.
- [35] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, 1976.
- [36] G. Shafer, Belief functions and parametric models, *J. Royal Statist. Soc. Ser. B* 44 (3) (1982) 322–352.
- [37] P. Smets, The transferable belief model for quantified belief representation, in: D.M. Gabbay, P. Smets (Eds.), *Handbook of Defeasible Reasoning and Uncertainty Management System*, vol. 1, Kluwer, Dordrecht, 1998, pp. 267–301.
- [38] J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior*, second edition, Princeton University Press, Princeton, NJ, 1947.
- [39] A. Wald, *Statistical Decision Function*, Wiley, New York, 1950.
- [40] P. Walley, Belief function representation of statistical evidence, *Ann. Statist.* 15 (4) (1987) 1439–1465.
- [41] L. Zadeh, Fuzzy set as a basis for a theory of possibility, *Fuzzy Sets and Systems* 1 (1978) 3–28.