Appeared in: Artificial Intelligence, Vol. 165, No. 2, 2005, pp. 137-163.

SCHOOL OF BUSINESS WORKING PAPER NO. 311

Decision Making on the Sole Basis of Statistical Likelihood

Phan H. Giang
Computer Aided Diagnosis & Therapy Group
Siemens Medical Solutions
51 Valley Stream Pkwy
Malvern, PA 19355
phan.giang@siemens.com

Prakash P. Shenoy
University of Kansas School of Business
1300 Sunnyside Ave, Summerfield Hall
Lawrence, KS 66045-7875 USA
pshenoy@ku.edu

November 2004[†]

[†] To appear in Artificial Intelligence journal.

Decision Making on the Sole Basis of Statistical Likelihood

Phan H. Giang
Computer Aided Diagnosis and Therapy Solutions (CAD)
Siemens Medical Solutions
51 Valley Stream Pkwy
Malvern, PA 19355 USA
phan.giang@siemens.com

Prakash P. Shenoy
University of Kansas School of Business
1300 Sunnyside Ave, Summerfield Hall
Lawrence, KS 66045-7585, USA
pshenoy@ku.edu

November 17, 2004

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.

Keywords: decision theory; choice under uncertainty; likelihood; statistical inference, ambiguity attitude.

1 Introduction

Various formal decision theories for choice under risk and uncertainty have been studied since the seminal work by von Neumann-Morgenstern (vNM) [37] 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. As the value of an axiomatic model is based on the acceptability of its assumptions, the debate on the value of the vNM theory started almost immediately with their publication [2]. As the result of this ongoing debate, axiomatic systems that are weaker than vNM but still possess the expected utility representation have been investigated [16, 32, 31]. 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 $\mathcal{F} = \{P_\theta | \theta \in \Omega\}$ 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 [8, 5, 4], 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 [38], 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.

[1] A preliminary version of this work has appeared in the Proceedings of 18th Conference on Uncertainty in Artificial Intelligence (UAI 2002) [20].

In this paper, we proposes 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.

This paper is organized as follows. In the next section, we discuss extending likelihood functions to an uncertainty measure—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. They are justified by intuition as well as by stochastic dominance principle. Next, we introduce the concept of binary utility and prove the expected utility theorem for likelihood lotteries. That is followed by comments on related works. In section 4, we apply our decision theory for a problem of statistical inference. Finally, 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 [17]. Fisher used likelihoods to measure "mental confidence" in competing scientific hypotheses as a result of a statistical experiment (see [13] 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 quan-

[2]One can write $P_{\theta_0}(y)$ in the form of a conditional probability: $P(Y = y|\theta = \theta_0)$. 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.

tities 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 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 [34] 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*)

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

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

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

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

Shafer [34] 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 $\Omega_i = \{\omega | lik_x(\omega) = a_i\}$ with $a_1 > a_2 \ldots > a_k$. Then there are k foci (A_i) and k masses

$$A_i = \cup_{j=1}^{i} \Omega_i \qquad \text{and} \qquad m(A_i) = \frac{(a_i - a_{i+1})}{a_1} \tag{3}$$

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

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

$$Pl(A \cup B) = \max(Pl(A), Pl(B)) \tag{4}$$

A crucial argument in favor of Shafer's original[3] 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 [29] by Neyman and Pearson (1928). ML is not only intuitively appealing, but it is also backed by various asymptotic optimality properties [25, 26].

We will use different notation than one in [34]. 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 $Lik_y : 2^{\Omega} \rightarrow [0, 1]$ as follows.

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

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

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

[3]Shafer [35] 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 [39] 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.

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 extended likelihood function by the following equation.

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

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 want to calculate the extended likelihood function 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

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

Table 1: Likelihood, extended likelihood and conditioning

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 $Lik_{1.4}(A|D) = Lik_{1.4}(\omega_2)/Lik1.4(D) = 0.6704$ 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 [40, 12]. What distinguishes a possibility function from a CBF is its fuzzy set semantics and, consequently, the notion of *ordinal conditioning*:

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

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.

3 A Decision Theory with Likelihood Uncertainty

Let us formalize the decision problem we will study. We assume a decision situation that includes a set Ω of simple hypotheses (parameter space); a set X of consequences/rewards and an observation y. The uncertainty about hypotheses is expressed by an extended likelihood function π calculated from y. An action a is a mapping $\Omega \to X$ i.e., the consequence of an action is determined by which hypothesis is true. It should be noted that the states on which rewards are dependent on are not observable. The set of actions is denoted by $\mathcal{A}$. For the sake of clarity, we assume that X is finite and its elements are denoted by $x_1, x_2, \ldots x_r$.

[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.

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 $[L(x_1)/x_1, L(x_2)/x_2, \ldots]$ 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 $[L_i/x_i]_{i=1}^m$, $\cup_{1 \leq i \leq m} a^{-1}(x_i) = \Omega$. Since π is an extended likelihood function 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 [28]). 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 $x_1 \succeq x_2 \succeq \ldots \succeq x_r$ 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, $\overline{x} \equiv x_1$ and $\underline{x} \equiv x_r$. A lottery that involves only the best $\overline{x}$ 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 $L = [\delta_1/L_1, \delta_2/L_2 \ldots \delta_k/L_k]$ and $L_i = [\kappa_{i1}/x_1, \kappa_{i2}/x_2, \ldots \kappa_{ir}/x_r]$ then $L \sim [\kappa_1/x_1, \kappa_2/x_2, \ldots \kappa_r/x_r]$ with $\kappa_j = \max_{1 \leq i \leq k}\{\delta_i.\kappa_{ij}\}$

(A3) Substitutability.
If $L_i \sim L_i'$ then $[\delta_1/L_1, \ldots \delta_i/L_i \ldots \delta_k/L_k] \sim [\delta_1/L_1, \ldots \delta_i/L_i' \ldots \delta_k/L_k]$

(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.

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

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 $L = [\delta_1/L_1, \delta_2/L_2, \ldots, \delta_k/L_k]$ where $L_i = [\kappa_{i1}/x_1, \ldots, \kappa_{ir}/x_r]$ 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 $\gamma_1, \gamma_2, \ldots \gamma_k$ with likelihoods $\delta_1, \delta_2, \ldots \delta_k$ 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 $\sigma_{o_i(1)}, \sigma_{o_i(2)}, \ldots \sigma_{o_i(r)}$ with likelihoods $\kappa_{i1}, \kappa_{i2}, \ldots \kappa_{ir}$ where o_i is a permutation of $(1, 2, 3, \ldots r)$. 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 $< \gamma_i \sigma_{o_i(j)} >$ is the true value of θ for $1 \leq i \leq k$. Because of conditioning equation 7, we have

$$Lik(\gamma_i \sigma_{o_i(j)}) = Lik(\gamma_i) Lik(\sigma = \sigma_{o_i(j)} | \gamma = \gamma_i) = \delta_i . \kappa_{ij} \tag{10}$$

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

$$Lik(\{\gamma_i \sigma_{o_i(j)} | 1 \leq i \leq k\}) = \max\{\delta_i . \kappa_{ij} | 1 \leq i \leq k\} \tag{11}$$

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. Figure 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 $c = [\lambda_1/\overline{x}, \lambda_2/\underline{x}]$ such that $x \sim c$. For clarity, let us assume that $\overline{x} = 1$, $\underline{x} = 0$ (the argument remains valid for any real values of $\overline{x}$ and $\underline{x}$ as long as $\overline{x} > \underline{x}$). For any $x \in [0, 1]$, 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 2 values $\{\theta_1, \theta_2\}$ that she will use for the unknown parameter. Probability distribution f_θ as well as

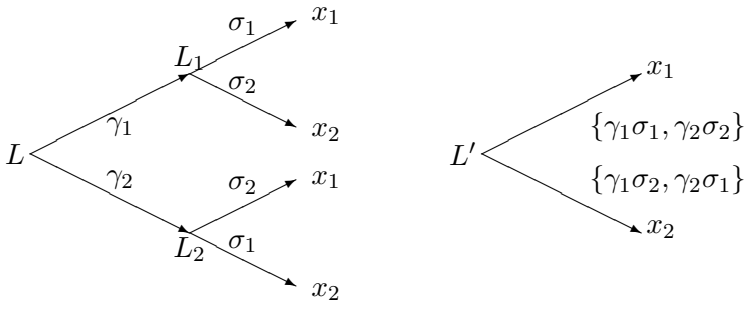

Figure 1: Two-stage L and one-stage L' are equivalent

the possible values of the parameter are revealed to all parties. For example, a normal distribution[5] with given standard devision $\sigma = 1$ 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 $\overline{x}$ 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,

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

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 $[Lik_{y_i}(\theta_1)/\overline{x}, Lik_{y_i}(\theta_2)/\underline{x}]$. 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 $[Lik_y(\theta_1)/\overline{x}, Lik_y(\theta_2)/\underline{x}]$.

Comparing the likelihood gamble with a probabilistic gamble[6] 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 dependent, not on an observation (y) but on the unobservable true value of

[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 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 Figure 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, observation $y = -3$ lends more support to hypothesis $\theta = -1$ than observation $y = 1$. 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.

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

The expected payoff of the likelihood gamble

$$\begin{aligned} V_y(\rho) &= P_\rho(\theta = \theta_1 | y) \cdot \overline{x} + P_\rho(\theta = \theta_2 | y) \cdot \underline{x} & (14) \\ &= \underline{x} + P_\rho(\theta = \theta_1 | y) \cdot (\overline{x} - \underline{x}) & (15) \end{aligned}$$

With $\overline{x} - \underline{x} > 0$, $V_y(\rho)$ is a strictly increasing function of $P_\rho(\theta = \theta_1 | y)$. When $\overline{x} = 1$ and $\underline{x} = 0$, Eq. 15 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 [27]. 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 $E(.)$ is the expectation operator.

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

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

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

Proof: 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

$$\frac{Lik_y(\theta_1)}{Lik_y(\theta_2)} > \frac{Lik_{y'}(\theta_1)}{Lik_{y'}(\theta_2)} \quad (17)$$

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. 17 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. 17 holds iff $Lik_y(\theta_2) < Lik_{y'}(\theta_2)$; (b) If $Lik_y(\theta_2) = Lik_{y'}(\theta_2) = 1$, Eq. 17 holds iff $Lik_y(\theta_1) > Lik_{y'}(\theta_1)$; (c) If $Lik_y(\theta_1) = Lik_{y'}(\theta_2) = 1$, Eq. 17 holds iff either $Lik_y(\theta_2) < 1$ or $Lik_{y'}(\theta_1) < 1$. By Eq. 9, we have Eq. 17 holds iff $[Lik_y(\theta_1)/1, Lik_y(\theta_2)/0] \succ [Lik_{y'}(\theta_1)/1, Lik_{y'}(\theta_2)/0]$. ■

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*

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

Proof: ($\Rightarrow$): 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

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

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 Figure 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.

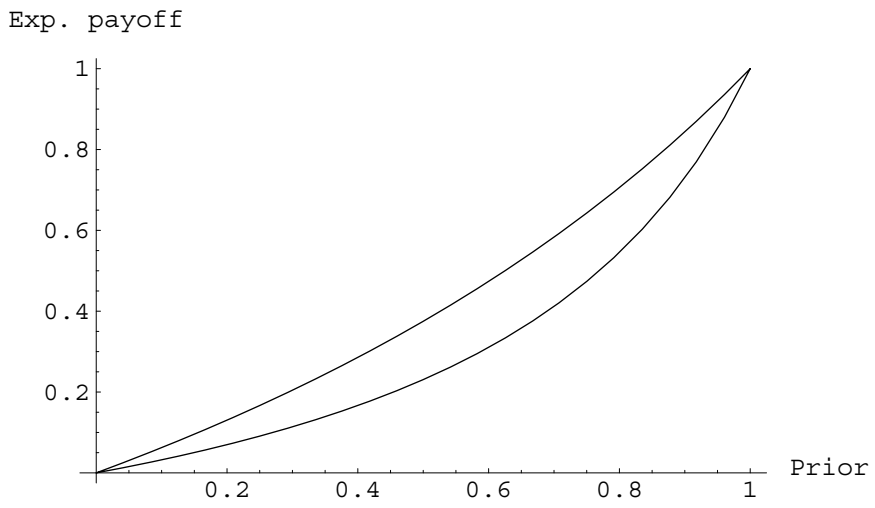

Figure 2: Expected payoff functions for $y = .60$ (lower curve) and $y = .26$ (upper curve)

3.2 Binary Utility

We will now proceed to study the preference relation satisfying the axioms. The following lemma shows that the set of lotteries divided by the indifference relation ($\mathcal{L}/\sim$) 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.*

Proof: 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 $x_i \sim s_i = [\kappa_{i1}/\overline{x}, \kappa_{ir}/\underline{x}]$ 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 $L = [\pi_1/x_1, \pi_2/x_2, \ldots, \pi_r/x_r]$, by A3, $L \sim L_1$ where $L_1 = [\pi_1/s_1, \pi_2/s_2, \ldots, \pi_r/s_r]$. 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 $[\kappa_{i1}/\overline{x}, \kappa_{i2}/x_2, \ldots, \kappa_{ir}/\underline{x}]$ with $\kappa_{ij} = 0$ for $2 \leq j \leq r-1$. By A2, $L_1 \sim L_2$ where $L_2 = [\kappa_1/\overline{x}, \kappa_2/x_2, \ldots, \kappa_r/\underline{x}]$ with $\kappa_j = \max\{\pi_i.\kappa_{ij} | 1 \leq i \leq k\}$. 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 $s_1 = s_2$. ■

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

$$\mathcal{U} \stackrel{\text{def}}{=} \{\langle a, b\rangle | a, b \in [0,1] \text{ and } \max(a,b) = 1\}. \tag{18}$$

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 $\gtrdot$ on $\mathcal{U}$ (to distinguish from the order $\geq$ on scalars) is defined as

$$\langle a, b\rangle \gtrdot \langle a', b'\rangle \quad \text{iff} \quad \begin{cases} a = a' = 1 \ \& \ b \leq b', \text{ or} \\ a = 1 \ \& \ a' \leq 1, \text{ or} \\ a \geq a' \ \& \ b = b' = 1 \end{cases} \tag{19}$$

Strict preference ($\gg$) and indifference ($=$) derivatives are also used. The special structure of $\mathcal{U}$ allows a simplification of order definition given in Eq. 19. The proof of the following lemma is straightforward and is therefore omitted.

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

$$\langle a, b\rangle \gg \langle a', b'\rangle \quad \textit{iff} \quad (a > a') \vee (b < b') \tag{20}$$

$$\langle a, b\rangle \gtrdot \langle a', b'\rangle \quad \textit{iff} \quad (a \geq a') \wedge (b \leq b') \tag{21}$$

$$\langle a, b\rangle = \langle a', b'\rangle \quad \textit{iff} \quad (a = a') \wedge (b = b') \tag{22}$$

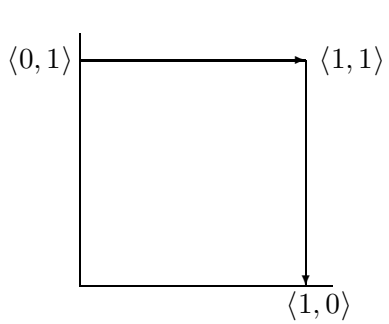

Figure 3: Binary utility scale $\mathcal{U}$.

We refer to $\mathcal{U}$ equipped with order $\geqq$ 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 [15]. 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.

To operate on binary utilities, we extend[7] product and max operations to work with pairs as follows. For scalar α, β, π and γ

$$\alpha.\langle\beta,\gamma\rangle) \stackrel{\text{def}}{=} \langle\alpha.\beta, \alpha.\gamma\rangle \tag{23}$$

$$\max(\langle\alpha,\beta\rangle, \langle\gamma,\pi\rangle) \stackrel{\text{def}}{=} \langle\, \max(\alpha,\gamma), \max(\beta,\pi)\rangle \tag{24}$$

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,* $\max(u, v) \geqq \max(u, v')$ *if* $v \geq v'$.

3.3 Representation Theorem

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

[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.

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

$$QU([\pi_1/L_1, \ldots, \pi_k/L_k]) = \max_{1\le i\le k} \{\pi_i.QU(L_i)\} \tag{25}$$

Proof:
($\Rightarrow$) Suppose $\succeq$ satisfies the axioms. We will show that there exists a function $QU : \mathcal{L} \to \mathcal{U}$ that satisfies Eq. 25 and represents $\succeq$. We construct function QU as follows. For a canonical lottery, define $QU([\lambda/\overline{x}, \mu/\underline{x}]) = \langle \lambda, \mu\rangle$. 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 19, for canonical lotteries s, s' we have $s \succeq s'$ iff $QU(s) \gtrdot QU(s')$. 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. 25. Consider depth-one lottery $L = [\pi_1/x_1, \ldots, \pi_r/x_r]$. By A4, each consequence x_i is indifferent to a canonical lottery, say $s_i = [\kappa_{i1}/\overline{x}, \kappa_{ir}/\underline{x}]$. Therefore, $QU(x_i) = \langle\, \kappa_{i1}, \kappa_{ir}\,\rangle$. Consider lottery $L' = [\pi_1/s_1, \pi_2/s_2, \ldots, \pi_r/s_r]$. From A3, $L \sim L'$. Using A2 and the argument in the proof of Lemma 3, we have L' is indifferent to canonical lottery $s = [\kappa_1/\overline{x}, \kappa_r/\underline{x}]$ where

$$\kappa_l = \max_{1\le i\le r} \{\pi_i.\kappa_{il}\} \quad \text{where} \quad l \in \{1, r\} \tag{26}$$

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

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

Thus, we show $QU(L) = \max_{1\le i\le r}\{\pi_i.QU(x_i)\}$. 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. 25 and $\langle 1,0\rangle, \langle 1,0\rangle \in QU(X)$ *i.e.,* $L \succeq_q L'$ iff $QU(L) \gtrdot QU(L')$. We show that $\succeq_q$ satisfies axioms A1 through A5. A1 is satisfied because relation $\gtrdot$ defined on $\mathcal{U}$ by Eq. 19 is reflexive, complete and transitive.
Let $L = [\pi_1/L_1, \ldots, \pi_i/L_i \ldots, \pi_k/L_k]$ and $L' = [\pi_1/L_1, \ldots, \pi_i/L'_i, \ldots, \pi_k/L_k]$. Assume $L_i \sim_q L'_i$. By definition of $\succeq_q$, it means $QU(L_i) = QU(L'_i)$. Apply Eq. 25 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 $L = [\pi_1/L_1, \pi_2/L_2, \ldots, \pi_k/L_k]$ where $L_i = [\kappa_{i1}/x_1, \kappa_{i2}/x_2, \ldots, \kappa_{ir}/x_r]$ 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. 25 for L_i and then L,

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

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

$$\kappa_j = \max_{1\leq i\leq k}\{\pi_j.\kappa_{ij}\}$$

Apply Eq. 25 for L_s, we have

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

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 $\overline{x}, \underline{x}$ the elements of X such that $QU(\overline{x}) = \langle 1, 0\rangle$ and $QU(\underline{x}) = \langle 0, 1\rangle$. By Eq. 19 and definition of $\succeq_q$, we have $\overline{x} \succeq_q x$ and $x \succeq_q \underline{x}$ for all $x \in X$. For any canonical lottery $[\lambda/\overline{x}, \mu/\underline{x}]$ where $\max\{\lambda, \mu\} = 1$, by Eq. 25 we have $QU([\lambda/\overline{x}, \mu/\underline{x}]) = \max\{\lambda.\langle 1, 0\rangle, \mu.\langle 0, 1\rangle\} = \max\{\langle \lambda, 0\rangle, \langle 0, \mu\rangle\} = \langle \lambda, \mu\rangle$. Thus, by Eq. 19 we conclude that A5 is satisfied.

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

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.

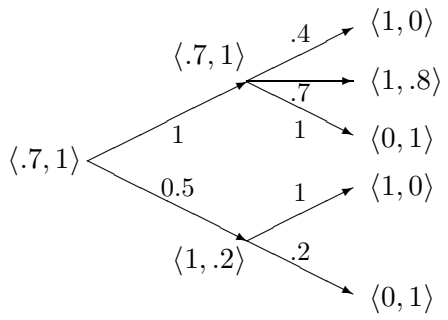

Figure 4: Qualitative utility calculation.

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. 25 can be rewritten as

$$QU([\pi_1/x_1, \ldots, \pi_r/x_r]) = \max_{1 \leq i \leq r} \{\pi_i . QU(x_i)\} \tag{27}$$

The form of QU in (27) resembles the vNM expected utility expression. The expected utility of a probabilistic lottery p on X is defined as $U(p) \stackrel{\text{def}}{=} \sum_{1 \leq i \leq r} p(x_i).U(x_i)$.

Maximization in (27) 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.

Figure 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 Archimedian properties [24].

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* $L_2 \succ [\lambda'/L_1, \mu'/L_3]$.

Proof: 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 $QU([\lambda/L_1, \mu/L_3]) \gg QU([\lambda/L_2, \mu/L_3])$. 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 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 $L_2 \succ [\lambda'/L_1, \mu'/L_3]$.■

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 [36, 7, 23]. 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 [36] 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 [14], this model produces an unintuitive solution. Halpern [23] 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 [19], 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]. As noted earlier, a possibility function satisfies equations (5, 6) that define an extended likelihood function. 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 [21]. 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 $Lik_y(\theta_1), Lik_y(1)$ into Eq. 13 and solve for ρ we find

$$\rho = \frac{x\mu}{(1-x)\lambda + x\mu} \tag{28}$$

We call ρ calculated by Eq. 28 an *implicit prior*[8] for the obvious reason.

[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 [18].

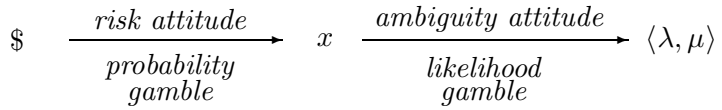

Figure 5: Risk and ambiguity in transformation from money to binary utility

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 $\overline{x}$ 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 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. However, risk attitude could be easily incorporated in our framework by assuming that elements of X in utility unit i.e., $x = u(d)$ where d has dollar unit and u is a dollar-to-utility function. Figure 5 illustrates the roles of risk and ambiguity attitudes in the transformation from monetary unit to binary utility via unary utility.

In Economics and Statistics literature, a prominent axiomatic decision theory without (additive) probability has been studied by Schmeidler [32] and extended by Gilboa [22] and Sarin and Wakker [31]. Schmeidler considers a preference relation $\succeq$ on acts (functions from Ω to X) and assumes it

[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., $\pi \succeq_{pes} \pi'$ ($\pi' \succeq_{opt} \pi$) if $\pi' \geq \pi$.

satisfies following axioms. Suppose f, g, h are acts.

S1 Weak order. $f \succeq g$ or $g \succeq f$; If $f \succeq g$ and $g \succeq h$ then $f \succeq h$.

S2 Co-monotonic independence. If f, g, h are pairwise co-monotonic and $\alpha \in [0,1]$, $f \succ g$ implies $\alpha f + (1-\alpha)h \succ \alpha g + (1-\alpha)h$.

S3 Continuity. If $f \succ g$ and $g \succ h$ then there are $\alpha, \beta \in]0,1[$ such that $\alpha f + (1-\alpha)h \succ g$ and $g \succ \beta f + (1-\beta)h$

S4 Monotonicity. If $f(s) \succeq g(s)$ for all $s \in \Omega$ then $f \succeq g$

S5 Non-degeneracy. Not for all f and g $f \succeq g$.

Two acts f and g are co-monotonic if for no two states s_1, s_2 $f(s_1) \succ f(s_2)$ and $g(s_2) \succ g(s_1)$. Schmeidler proves a representation theorem: $\succeq$ satisfies the axioms $S1 - S5$ iff there is a unique capacity measure v on 2^{Ω} and

$$f \succeq g \text{ iff } CEU_v(f) \geq CEU_v(g)$$

A real-valued function $v : 2^{\Omega} \to [0,1]$ is called *capacity* function if $v(\emptyset) = 0$, $v(\Omega) = 1$ and for $A \subseteq B \subseteq \Omega$ $v(A) \leq v(B)$. $CEU_v(f)$ - *Choquet expected utility* of act f with respect to capacity v - is defined as follows. For simplicity, we assume that X is a set of reals, interpreted as utilities, and is ordered $x_0 > x_1 > \ldots > x_m$. For an act f, $A_k^f \stackrel{\text{def}}{=} \{s \in \Omega | f(s) \geq x_k\}$ - set of states where f delivers x_k or better consequence. Obviously A_i^f are nested i.e., $A_0^f \subseteq A_1^f \subseteq \ldots \subseteq A_m^f = \Omega$

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

The major difference from von Neumann-Morgenstern's approach is in axiom $S2$ that stipulates that independence is applicable for co-monotonic acts only.

Since likelihood functions satisfy the requirements for a capacity function, it makes sense to compare our approach with Schmeidler's. Notice also that both approaches have representation theorems. Therefore, the discussion could be either in terms of the axioms or in terms of the utility functions.

First, *CEU* is a real-valued function while *QU* is not. In our setting, consequences are bounded (by $\overline{x}$ and $\underline{x}$). Schmeidler's setting has no such restriction. It is easy to show that preference relation $\succeq_{QU}$, represented by *QU*, satisfies $S1$, $S4$ and $S5$ but not $S2$ and $S3$.

Let us examine the preference relation $\succeq_{CEU}$, represented by *CEU*, in relation to our axioms $A1 - A5$. Clearly, since $A1$ is the same as $S1$, $\succeq_{CEU}$ satisfies $A1$.

Strictly speaking, likelihood functions, which satisfy requirements for a capacity measure, are not a special case of capacity function. The most important difference is that while conditioning is well defined for likelihood function, it is not clear if and how such an operation could be defined for capacity functions. Suppose (for the sake of an argument) that such an operation exist. Then reduction of compound lotteries axiom $A2$ is a reasonable requirement. However, it is unlikely that would be the case. Here is the evidence. In Schmeidler's approach, each act is a two-stage lottery. The first stage is a *horse* lottery whose uncertainty is described by a capacity function. The rewards of a horse lottery are *roulette* lotteries. The uncertainty pertaining to the roulette lotteries is described by a probability function. In [31], Sarin and Wakker consider one-stage setting where the event algebra for lotteries contains a sub-algebra of "unambiguous" events for which uncertainty is probability. They are able to retain Choquet expected utility representation. Sarin and Wakker show [31] that their approach and Schmeidler's one are irreconcilable. This result implies that preference relation represented by *CEU* does not satisfy property like $A2$.

$\succeq_{CEU}$ also violates qualitative monotonicity axiom ($A5$). Let us calculate *CEU* for canonical acts with respect to likelihood function π assuming for simplicity $\overline{x} = 1$ and $\underline{x} = -1$. For a canonical act a that delivers the best consequence ($\overline{x}$) if A occurs and the worst consequence ($\underline{x}$) otherwise ($\overline{A}$ occurs). This means $A_0^a = A$ and $A_1^a = \Omega$.

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

First we observe that *CEU* for a canonical act does not depend on the capacity of event that leads to worst consequence. $v(\overline{A}) \neq 1 - v(A)$ because a capacity measure is non-additive. In the case of likelihood function π, equality $\pi(A) = 1$ does not imply any information about the value of $\pi(\overline{A})$. *CEU* equalizes canonical acts which have the same capacity on the best consequence. $A5$ requires a comparison of capacities of getting worst consequence when the capacities of getting the best consequence are equal.

Schmeidler's *CEU* representation is the result of considering (i) separation of "utility" from "probability" and (ii) "functional representations which are the sum of products of two numbers; one number has a "probability" interpretation and the other has a "utility" interpretation" ([32] p.584). The resulting capacity function is implicit and its updating operation is not

explicitly considered. 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".

This observation seems to suggest that *CEU* is not appropriate for existing (non-probabilistic) uncertainty calculi such as Dempster-Shafer belief functions [34], fuzzy possibility [40] and plausibility measures [23] 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 $v'(A) \stackrel{\text{def}}{=} 1 - v(\overline{A})\ \forall A \subseteq \Omega$. 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 (*Bel*) and plausibility (*Pl*) functions in Dempster-Shafer theory. Thus, from a decision making point of view, 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 $\delta : \mathcal{Y} \to \mathcal{A}$, that is for an observation y the rule recommends an action $\delta(y)$. The *risk function* of a decision rule δ at parameter value θ is defined as

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

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

[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.

of the risk function. In case no a priori information about parameter exists, Wald [38] advocated the use of minimax rule which minimizes the worst risk that could be attained by a rule.

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

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*

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

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

$$\delta^*_{Bayes,\rho} = \arg\min_{\delta\in\Delta} r(\delta) \tag{32}$$

Wald [38] 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

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

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

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

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., $r(\delta^*_P) = r(\delta^*_{Bayes,\rho})$*

Proof: We need to show that the Bayes risk for δ^*_P is minimal i.e.,

$$\forall\delta \in \Delta\ r(\delta^*_P) \le r(\delta) \tag{35}$$

Substitute Eq. 29 into Eq. 31, we have

$$r(\delta) = \int_\Omega\int_\mathcal{Y} V(\delta(y),\theta)p_\theta(y)\rho(\theta) = \int_\mathcal{Y}\int_\Omega V(\delta(y),\theta)p_\theta(y)\rho(\theta) \tag{36}$$

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. ■

4.2 Likelihood solution

Without knowing prior ρ, we propose the following solution based on the logic that leads to δ_P^*. For each $y \in \mathcal{Y}$, there is an associated extended likelihood function Lik_y. An action together with a likelihood function induce a likelihood lottery. For an action a and an observation y, denote by $L_a(y)$ the lottery that is generated. Likelihood lotteries are compared by QU. The selected action given observation y is

$$a^{Lik}(y) = \arg\sup_{a\in\mathcal{A}} QU(L_a(y)) \tag{37}$$

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 one concerns information. How much information is assumed to be available to the decision maker and how it is utilized. In order to apply the Bayes solution, we must know the prior distribution of the unknown parameter. As mentioned earlier, in many situations, such an assumption is not realistic. However, if the prior is known, the posterior probability could be calculated, and the Bayes solution makes full use of this extra-experiment information as well as the information provided by experimental results (likelihood). In Wald's proposal, the risk function has no *special* role for the actually observed data, thus, it ignores information about the parameter obtainable from the data. One can argue that the minimax rule reflects a cautious attitude. But it is, in our opinion, too cautious. 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

[11] In contrast to *max* defined in Eq. 24 that operates on scalars $\geq$.

r.v. are non-negative, i.e., cumulative distribution functions satisfy $F(0) = 0$. For cumulative distribution function F, define for any natural n

$$F_n(z) = \int_0^z F_{n-1}(x)dx \tag{38}$$

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.

As mentioned in the introduction, statistical inference, via its manifestation as probabilistic model selection, should be of interest for researchers in AI, machine learning, pattern recognition and data mining. However, most works in these areas continue to select a model using simple criteria such as maximum likelihood (ML) or maximum a posteriori probability (MAP). From a decision theoretic point of view, the use of these criteria is equivalent to assuming equal utilities (costs) for all models under consideration. Clearly, it is a gross simplification. For example, between 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 (principle of parsimony). In the statistics literature, models are selected by using Akaike Information Criterion (AIC) [1] or Schwarz's criterion (also known as Bayesian Information Criterion or BIC) [33]. 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 [30] 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 that can be drawn from these works is that different models are associated with *different* costs. And therefore, these costs must be taken into account in a model selection process.

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 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. The symptom can be viewed as observation. 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.

	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

In this case, the Bayes solution is δ_5. Since the Bayes solution depends on prior $p(\theta_1)$, a sensitivity analysis shows Bayes solutions for different values of prior.

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

In our approach, suppose monetary values are translated into binary utilities 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. 28 we have

$$0.5 = \left(\frac{(1 - 0.8)\lambda}{0.8\mu} + 1 \right)^{-1} \tag{39}$$

From that we find $\lambda/\mu = 4$. 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 observation y_1, the likelihood function is $Lik_{y_1}(\theta_1) = .14$ and $Lik_{y_1}(\theta_2) = 1$. Action a_1 corresponds to lottery $L_{a_1}(y_1) = [.14/\langle 1, .25\rangle, 1/\langle .43, 1\rangle]$ whose qualitative expected utility is

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

Action a_2 is associated with lottery $L_{a_2}(y_1) = [.14/\langle 1, 1\rangle, 1/\langle 1, 1\rangle]$ whose qualitative expected utility $QU(L_{a_2}(y_1)) = \langle 1, 1\rangle$. Thus, given y_1, we have $a_2 \succ_{y_1} a_1$ i.e., a_2 is strictly preferred to a_1. Given observation y_2, the extended likelihood function is $Lik_{y_2}(\theta_1) = 1$ and $Lik_{y_2}(\theta_2) = .5$. We calculate qualitative expected utility for a_1 is $QU(L_{a_1}(y_2)) = \langle 1, .5\rangle$ and for a_2 $QU(L_{a_2}(y_2)) = \langle\, 1, 1\,\rangle$. Thus, $a_1 \succ_{y_2} a_2$. Given observation y_3, the extended likelihood function is $Lik_{y_2}(\theta_1) = 1$ and $Lik_{y_2}(\theta_2) = .2$. Qualitatively expected utility for a_1 is $QU(L_{a_1}(y_3)) = \langle 1, .25\rangle$ and for a_2 remains $QU(L_{a_2}(y_3)) = \langle 1, 1\rangle$. Thus, $a_1 \succ_{y_3} a_2$. In summary, our approach suggests δ_5 as the likelihood solution.

Let us make an informal comparison of likelihood solution with minimax and Bayes solutions. In this example, likelihood solution δ_5 while the minimax solution is δ_8. It is because, as we noted, minimax solution ignores the uncertainty generated by an observation while likelihood solution does not. In this sense, likelihood solution is more information efficient.

If the prior probability $p(\theta_1) = .7$, then the Bayes solution is δ_5 the same as the likelihood solution. If prior probability is available, one can argue that Bayes solution is *the* optimal one. However, the "optimality" of the Bayes solution does not come without cost. The requirement of prior probability can be satisfied either at some monetary cost (doing research, or buying from those who have). Alternatively, the decision maker can just assume in an ad hoc manner some prior distribution. This however would compromise the claimed optimality of a Bayes solution. 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.

One should be careful not to draw too much from the coincidence of the likelihood solution (δ_5) and the Bayes solution that corresponds to the largest prior interval. It is a result of several factors some of those are ad

hoc (e.g., unary-binary utility conversion). However, as we pointed out, axioms $A1$ to $A5$ on which likelihood solution is based, are structurally similar to those used by Luce and Raiffa [28] to justify the expected utility maximization principle which ultimately is the basis for Bayes solutions. Thus, at the 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 has precedence over the other depends on how much information is available.

5 Summary and Conclusion

In this paper, we develop a decision theory that utilizes likelihood information without assuming the existence of prior probability. We extend likelihood function as the uncertainty measure pertaining to the statistical inference problem. The extension, conforming to the practice of 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 used by von Neumann-Morgenstern for 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 the 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/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.

We are investigating potential applications of the results of this work for the problem of probabilistic model selection.

References

[1] Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information*

Theory (Budapest, 1973), B. N. Petrox and F. Caski, Eds., Academiai Kiado, p. 267.

[2] Allais, M. The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School. In *Expected Utility Hypotheses and the Allais Paradox*, M. Allais and O. Hagen, Eds. D. Reidel, Dordrecht, Boston, London, 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, Vol. XL, Paris, 1953 pp. 275-332.

[3] Barnett, V. *Comparative Statistical Inference*, 3 ed. John Wiley and Sons, New York, Chichester, Brisbane, 1999.

[4] Basu, D., and Ghosh, I. *Statistical Information and Likelihood: A Collection of Critical Essays by Dr. Basu*. Lecture notes in Statistics. Springer-Verlag, New York, Berlin, Heidelberg, 1988.

[5] Berger, J. O., and Wolpert, R. L. *The Likelihood Principle*, 2 ed. Lecture notes-Monograph series. Institute of Mathematical Statistics, Hayward, California, 1988.

[6] Borch, K. Utility and stochastic dominance. In *Expected utility hypotheses and the Allais paradox*, M. Allais and O. Hagen, Eds. D. Reidel, Dordrecht, Boston, London, 1979, pp. 193–202.

[7] Brafman, R. I., and Tennenholtz, M. An axiomatic treatment of three qualitative decision criteria. *Journal of the Association of Computing Machinary 47*, 3 (2000), 452–483.

[8] Cox, D. R., and Hinkley, D. V. *Theoretical Statistics*. Chapman and Hall, London, 1974.

[9] Dempster, A. A generalization of Bayesian inference. *Journal of Royal Statistical Society, Series B 30* (1968), 205–247. with discussion.

[10] Dubois, D., Godo, L., Prade, H., and Zapico, A. On the possibilistic decision model: from decision under uncertainty to case-based decision. *International Journal of Uncertainty, Fuzziness and Knowledge- based Systems 7*, 6 (1999), 631–670.

[11] Dubois, D., Moral, S., and Prade, H. A semantics for possibility theory based on likelihoods. *Journal of Mathematical analysis and applications 205* (1997), 359–380.

[12] Dubois, D., Nguyen, T. H., and Prade, H. Possibility theory, probability and fuzzy sets. In *Handbook of Fuzzy Sets Series*, D. Dubois and H. Prade, Eds. Kluwer Academic, Boston, 2000, pp. 344–438.

[13] Edwards, A. W. F. *Likelihood*. Cambridge Univesity Press, Cambridge, 1972.

[14] Ellsberg, D. Risk, ambiguity and the Savage's axioms. *Quarterly Journal of Economics 75*, 4 (1961), 643–669.

[15] Fishburn, P. C. Lexicographic orders, utilities and decision rules: A survey. *Management Scicence 20*, 11 (1974), 1442–1471.

[16] Fishburn, P. C. *Nonlinear Preference and Utility Theory*. The Johns Hopkins University Press, Baltimore, London, 1988.

[17] Fisher, R. A. On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A 222* (1922), 309–368. Reprinted: Collected Papers of R. A. Fisher vol. 1, ed. J. H. Bennett, University of Adelaide 1971.

[18] Giang, P. H. *A Decision Theory for Non-Probabilistic Uncertainty and Its Applications*. PhD thesis, University of Kansas, Lawrence, Kansas, 2003.

[19] Giang, P. H., and Shenoy, P. P. A qualitative linear utility theory for Spohn's theory of epistemic beliefs. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI–2000)* (San Francisco, CA, 2000), C. Boutilier and M. Goldszmidt, Eds., Morgan Kaufmann, pp. 220–229.

[20] Giang, P. H., and Shenoy, P. P. Statistical decisions using likelihood information without prior probabilities. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI–2002)* (San Francisco, CA, 2002), A. Darwiche and N. Friedman, Eds., Morgan Kaufmann, pp. 170–178.

[21] Giang, P. H., and Shenoy, P. P. Two axiomatic approaches to decision making using possibility theory. *European Journal of Operational Research 162*, 2 (2005), 450–467.

[22] Gilboa, I. Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics 16* (1987), 65–88.

[23] Halpern, J. Y. *Reasoning about Uncertainty*. MIT Press, Cambridge, Massachusetts, 2003.

[24] Jensen, N. E. An introduction to Bernoullian utility theory I: Utility functions. *Swedish Journal of Economics 69* (1967), 163–183.

[25] Kallenberg, W. C. M. *Asymtotic Optimality of Likelihood Ratio Tests in Exponential Families*, vol. 77 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1978.

[26] Lehmann, E. L., and Casella, G. *Theory of Point Estimation*, 2 ed. Springer Texts in Statistics. Springer, New York, Berlin, Heidelberg, 1998.

[27] Levy, H. Stochastic dominance and expected utility: Survey and analysis. *Management Science 38*, 4 (1992), 555–593.

[28] Luce, R. D., and Raiffa, H. *Games and Decision*. John Wiley & Sons, 1957.

[29] Neyman, J., and Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. part 1. *Biometrica 20A* (1928), 175–240. In J. Neyman and E. S. Pearson Joint Statistical Papers. University of California Press. 1967.

[30] Poland, W. B., and Shachter, R. D. Three approaches to probability model selection. In *Uncertainty in Artificial Intelligence: Proceedings of the Tenth Conference (UAI–94)* (1994), R. Lopez de Mantaras and D. Poole, Eds., Morgan Kaufmann.

[31] Sarin, R., and Wakker, P. A simple axiomatization of nonadditive expected utility. *Econometrica 60*, 6 (1992), 1255–1272.

[32] Schmeidler, D. Subjective probability and expected utility without additivity. *Econometrica 57*, 3 (1989), 571–587.

[33] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics 6* (1978), 461–464.

[34] Shafer, G. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[35] Shafer, G. Belief functions and parametric models. *Journal of the Royal Statistical Society, Series B 44*, 3 (1982), 322–352.

[36] Smets, P. The transferable belief model for quantified belief representation. In *Handbook of Defeasible Reasoning and Uncetainty Management system*, D. M. Gabbay and P. Smets, Eds., vol. 1. Kluwer, Doordrecht, 1998, pp. 267–301.

[37] von Neumann, J., and Morgenstern, O. *Theory of Games and Economic Behavior*, 2 ed. Princeton University Press, Princeton, NJ, 1947.

[38] Wald, A. *Statistical Decision Function*. John Wiley and Sons, New York, 1950.

[39] Walley, P. Belief function representation of statistical evidence. *The Annals of Statistics 15*, 4 (1987), 1439–1465.

[40] Zadeh, L. Fuzzy set as a basis for a theory of possibility. *Fuzzy Sets and Systems 1* (1978), 3–28.