On the Suboptimality of the Generalized Bayes Rule and Robust Bayesian Procedures from the Decision Theoretic Point of View: A Cautionary Note on Updating Imprecise Priors

THOMAS AUGUSTIN
Ludwig-Maximilians University of Munich, Germany

Abstract

This paper discusses fundamental aspects of inference with imprecise probabilities from the decision theoretic point of view. It is shown why the equivalence of prior risk and posterior loss, well known from classical Bayesian statistics, is no longer valid under imprecise priors. As a consequence, straightforward updating, as suggested by Walley's Generalized Bayes Rule or as usually done in the Robust Bayesian setting, may lead to suboptimal decision functions. As a result, it must be warned that, in the framework of imprecise probabilities, updating and optimal decision making do no longer coincide.

Keywords

decision making, generalized risk, generalized expected loss, imprecise prior risk and posterior loss, robust Bayesian analysis, Generalized Bayes rule

1 Introduction

A powerful method of inference has to provide answers to (at least) the following three questions:

- What is updating?
- How to learn from data? (inference)
- How to make optimal decisions?

The classical Bayesian statistical theory, based on precise probabilities, claims to provide a comprehensive framework to deal with all these aspects simultaneously. For a Bayesian, inference and decision making coincide, and the solution to both tasks is essentially based on updating prior probabilities by means of the Bayes rule. More precisely, Bayesian statistics is based on two paradigms [P1] and [P2], where

[P1] Every uncertainty can adequately be described by a classical probability distribution. This in particular allows to assign a prior distribution $\pi(\cdot)$ on parameter spaces in inferential problems and on the space of states of nature in decision problems.

[P2] After having observed the sample $\{x\}$, the posterior $\pi(\cdot|x)$ contains all the relevant information. Every inference procedure depends on $\pi(\cdot|x)$, and only on $\pi(\cdot|x)$.

There are several strong arguments for [P2], see, for instance, the discussion in [25]. Among them is the decision theoretic foundation by the often so-called 'main theorem of Bayesian decision theory': As discussed below, it says that decision functions with minimal risk under a prior $\pi(\cdot)$ can be constructed from considering optimal actions with respect to the posterior probability $\pi(\cdot|x)$ as an 'updated prior'.

In the last decade a rapidly increasing number of researches have objected against $[P1]$, and so theories of imprecise probabilities and interval probability emerged (see, e.g., the monographs by Walley [33], Kuznetsov [22], Weichselberger [39], the conference proceedings de Cooman, Fine, Moral and Seidenfeld [6] and the web page de Cooman and Walley [7]), offering a comprehensive framework to deal with a more realistic and reliable description of uncertainty. In this context also concepts generalizing conditional probability have been developed, suggesting the straightforward extension of $[P2]$, namely to use imprecise posteriors to update imprecise priors. This approach is discussed, among others, by Levi ([23],[24]), and is rigorously justified by general coherence axioms in Walley's theory ([33]). Moreover, it is even often understood as self-evident, and applied in many cases without a moment of hesitation, for instance, in the robust Bayesian Analysis (e.g., [35, 26]) and in economic applications following Kofler and Menges' [21] approach of decision making under linear partial information.[1]

The self-evidence of this way to proceed is questioned here. From a rigorous decision theoretic point of view, which is taken up in this paper, it is becoming clear without any ifs and buts that – quite surprisingly – such a procedure may be suboptimal: the resulting decision function may have higher risk than the optimal decision function. The present paper wants to illuminate this aspect. To achieve this goal, it proceeds as follows: Section 2 collects basic notions needed

[1] For further references see, e.g., Cozman's survey ([8]) on computational aspects and the references in [41, Section 1].

later from classical decision theory. After recalling some general aspects and terminology from the theory of interval probability in Section 3.1, both ingredients are melt together in Section 3.2, where the general framework for decision making under interval probability developed in [1, 2] is described briefly. Behind this background Section 4 explores the suboptimality of decision functions based on imprecise posteriors, while Section 5 returns to the fundamental questions formulated above and concludes with a short reflection on the consequences to be drawn from the observation made here.

2 Classical Decision Theory

2.1 The Basic Decision Problem and the Data Problem

Classical decision theory provides a formal framework for decision situations under uncertainty. The decision maker aims at choosing an *action* from of a nonempty, finite set $\mathrm{IA} = \{a_1, \ldots, a_i, \ldots, a_n\}$ of possible actions. Apart from trivial border cases, the consequences of every action depend on the true, but unknown *state* of nature $\vartheta \in \Theta = \{\vartheta_1, \ldots, \vartheta_j, \ldots, \vartheta_m\}$. The corresponding outcome is evaluated by a *loss function*

$$\begin{array}{rccc} l & : & (\mathrm{IA} \times \Theta) & \rightarrow & \mathbb{R} \\ & & (a, \vartheta) & \mapsto & l(a, \vartheta) \end{array}$$

and by the associated random variable $\mathbf{l}(a)$ on $(\Theta, \mathcal{P}o(\Theta))$ taking the values $l(a, \vartheta)$. For brevity of reference, the relevant components, the set IA of actions, the set Θ of states of nature and the precise loss function[2] $l(\cdot)$, is collected in the triple $(\mathrm{IA}, \Theta, l(\cdot))$, which is called *basic decision problem.*

For many applications it will prove of value to extend the problem by allowing for *randomized actions.* Formally, every randomized action can be identified with a classical probability $\lambda(\cdot)$ on $(\mathrm{IA}, \mathcal{P}o(\mathrm{IA}))$ where $\lambda(\{a\})$, $a \in \mathrm{IA}$, is interpreted as the probability to choose action a. The set of all randomized actions will be denoted by $\Lambda(\mathrm{IA})$. Pure actions, i.e. elements a of IA itself, are identified with the Dirac measure in the point $\{a\}$, and therefore are also understood to be elements of $\Lambda(\mathrm{IA})$. The loss function is extended to the domain $\Lambda(\mathrm{IA}) \times \Theta$ by $l(\lambda, \vartheta_j) := \sum_{i=1}^{n} \lambda(a_i) \cdot l(a_i, \vartheta_j)$. Analogously to $\mathbf{l}(a)$, $\mathbf{l}(\lambda)$ is that random variable which gives the loss of λ in dependence on the true state ϑ.

Quite often it is possible to obtain some information on the states of nature by collecting additional data. Formally, this can be described by an additional 'experiment' where the probability $p_\vartheta(\cdot)$ of the outcomes depends on the true state ϑ of nature. Let $\mathcal{X}$ be the sample space of this experiment, and assume

[2] Throughout the paper it is assumed that a (precise) loss function is given. On the *construction* of loss functions in the presence of ambiguity, generalizing the Neumann Morgenstern approach, see, e.g., [14] and the references therein.)

throughout the paper $\mathcal{X}$ to be finite, so that $\mathcal{X} = \{x_1, \ldots, x_s, \ldots, x_k\}$. The triple $(\mathcal{X}, \mathcal{P}o(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$ is called *sample information*, the basic decision problem together with the sample information *data problem*.

Now the decision problem consists in the choice between *decision functions (strategies)*

$$\begin{array}{rccc} d: & \{x_1, \ldots, x_k\} & \to & \Lambda(\mathrm{IA}) \\ & x & \mapsto & d(x) = \lambda\,, \end{array}$$

i.e. functions which map every observation x into a (randomized) action λ which has to be chosen if x occurs. Let $\mathbb{D}$ be the set of all decision functions. Decision functions are compared via their overall expected loss under $p_\vartheta(\cdot)$, i.e. one considers the so called *risk function*

$$R(d,\vartheta) := \sum_{s=1}^{k} l(d(x_s),\vartheta)\cdot p_\vartheta(x_s)\,, \tag{1}$$

which produces, analogous to above, the random variable $\mathbf{R}(d)$.

2.2 Optimality Criteria

If the states of nature are produced by a perfect random mechanism (e.g. an ideal lottery), and the corresponding probability measure $\pi(\cdot)$ on $(\Theta, \mathcal{P}o(\Theta))$ is completely known, the Bernoulli principle is nearly unanimously favored. One chooses that action λ^* which minimizes the expected loss

$$\mathbb{E}_\pi \mathbf{l}(\lambda) = \sum_{j=1}^{m} l(\lambda,\vartheta_j)\cdot\pi(\{\vartheta_j\}) \tag{2}$$

among all $\lambda \in \Lambda(\mathrm{IA})$, and that decision function which minimizes the expected risk

$$\mathbb{E}_\pi \mathbf{R}(d) = \sum_{j=1}^{m} R(d,\vartheta_j)\cdot\pi(\{\vartheta_j\}) \tag{3}$$

among all $d \in \mathbb{D}$, respectively.

In most practical applications, however, the true state of nature can not be understood as arising from an ideal random mechanism. And even if so, the corresponding probability distribution will be not known exactly. There are two main directions to proceed in this situation:

Since for a classical subjectivist, or Bayesian, according to $[P1]$, every situation under uncertainty can be described by a single, precise probability measure $\pi(\cdot)$, the lack of such a known random mechanism does not make any important difference to the decision maker. (S)he acts according to *subjective expected loss/risk*. In this context a special terminology became quite common: $\pi(\cdot)$ is called *prior probability*, and the expression in (3) *prior risk*.

In contrast, from the viewpoint of an 'objectivist' it does not make any sense at all to assign a probability on $(\Theta, \mathcal{P}o(\Theta))$. Therefore, the objectivist concludes that the decision maker is completely ignorant about which state of nature will occur; (s)he has to act according to a criterion based on complete ignorance. The most common criterion is the *minimax rule*, which concentrates on the worst state of nature, leading in the basic decision problem to

$$\max_{\vartheta\in\Theta} l(\lambda,\vartheta) \to \min \tag{4}$$

and in the data problem to

$$\max_{\vartheta\in\Theta} R(d,\vartheta) \to \min\ . \tag{5}$$

2.3 The Main Theorem of Bayesian Decision Theory

It is quite an essential characteristic of Bayesian decision theory that an optimal decision function $d^*(\cdot)$ minimizing the prior risk (3) can be obtained by minimizing, for every observation $\{x\}$, the *posterior loss*,

$$\mathbb{E}_{\pi(\cdot|x)}\mathbf{l}(\lambda) = \sum_{j=1}^{m} l(\lambda,\vartheta_j)\cdot\pi(\{\vartheta_j\}|x) \tag{6}$$

where, compared to (2), the prior $\pi(\cdot)$ is replaced by the 'updated prior', i.e., the posterior $\pi(\cdot|x)$. This is the decision theoretic foundation for the usual Bayesian updating (see also $[P2]$ from the Introduction). More precisely this fundamental relation is formulated in

Proposition 1 ("Main theorem of Bayesian decision theory") [3] *Consider a data problem, consisting of a basic decision problem* $(\mathrm{IA},\Theta,l(\cdot))$*, a sample information* $(\mathcal{X},\mathcal{P}o(\mathcal{X}),(p_\vartheta(\cdot))_{\vartheta\in\Theta})$ *and a prior probability* $\pi(\cdot)$*. For every* $s = 1,\ldots,k$*, let* $\pi(\cdot|x_s)$ *be the corresponding posterior given* x_s*, and* λ_s^* *be an optimal solution to the basic decision problem with respect to* $\pi(\cdot|x_s)$*, i.e. an action minimizing (6).*

Then $d^* := (\lambda_1^*,\ldots,\lambda_s^*,\ldots,\lambda_k^*)$ *is an optimal decision function minimizing (3).*

Remark 1 *The property formulated in Proposition 1 is constitutive for Bayesian decision making. In particular, an analogous reduction of the data problem to basic decision problems is not possible for the maximin criterion (4) and (5).*

3 Decision Making under Interval Probability

It has often been complained that both classical ways to proceed – relying on subjective expected loss as well as acting according to a criterion based on complete ignorance – are inappropriate, because they both distort the *partial* nature of

[3] Compare, for instance, [4, p. 159, Result 1].

the knowledge on the decision maker's hand: The objectivist's criteria treat partial knowledge like complete ignorance, often leading to unsatisfactory, overpessimistic solutions. Subjective utility/loss theory on the other hand identifies partial knowledge with complete probabilistic knowledge. This conflicts with Ellsberg's [11] experiments, which made it perfectly clear that ambiguity (i.e. the deviation from ideal stochasticity) plays a constitutive role in decision making — neglecting it may lead to deceptive conclusions.

Imprecise probabilities and related concepts are understood to provide a powerful language which is able to reflect the partial nature of the knowledge suitably and to express the amount of ambiguity adequately. (See [7] and [39, Ch. 1] for recent reviews on the development in this field.)

3.1 Basic Terminology of Interval Probability

With respect to the intended application the whole consideration is restricted here to the case of a finitely generated algebra $\mathcal{A}$ based on a sample space Ω. Then, without loss of generality, Ω is finite, and $\mathcal{A}$ is the power set of $\Omega = \{\omega_1, \ldots, \omega_k\}$.

To distinguish in terminology, every probability measure in the usual sense, i.e. every set function $p(\cdot)$ satisfying Kolmogorov's axioms is called a *classical probability*. The set of all classical probabilities on the measurable space $(\Omega, \mathcal{A})$ will be denoted by $\mathcal{K}(\Omega, \mathcal{A})$.

Axioms for interval-valued probabilities $P(\cdot) = [L(\cdot), U(\cdot)]$ can be obtained by looking at the relation between the non-additive set-function $L(\cdot)$ and $U(\cdot)$ and the set of classical probabilities being in accordance with them. On a finite sample space, as considered throughout this paper, several concepts of interval probability coincide. They all are concerned with set-functions

$$\begin{array}{rccl} P(\cdot) \;:\; \mathcal{A} & \to & \mathcal{Z}_0 := \{[L,U] \,|\, 0 \le L \le U \le 1\} \\ A & \mapsto & P(A) = [L(A), U(A)] \end{array}$$

with

$$\mathcal{M} := \{p(\cdot) \in \mathcal{K}(\Omega, \mathcal{A}) \mid L(A) \le p(A) \le U(A),\ \forall A \in \mathcal{A}\} \neq \emptyset\,. \tag{7}$$

and

$$\left.\begin{array}{l} \inf\limits_{p(\cdot)\in\mathcal{M}} p(A) = L(A) \\ \sup\limits_{p(\cdot)\in\mathcal{M}} p(A) = U(A) \end{array}\right\} \quad \forall A \in \mathcal{A}\,. \tag{8}$$

Such $P(\cdot)$, and the corresponding set functions $L(\cdot)$ and $U(\cdot)$, are called lower and upper probability ([17]), envelopes ([34, 9]), coherent probability ([33]) and F-probability ([37, 38, 39]). In the game theoretic setting $\mathcal{M}$ is the 'core'. Here Weichselberger's terminology is used calling $\mathcal{M}$ *structure*. Note that, by (8), there is a one-to-one correspondence between $P(\cdot)$ and the structure $\mathcal{M}$.

Two-monotone capacities ([17], also called supermodular capacities ([9]) or convex capacities ([18]), as well as belief functions ([28, 42]) are special cases. More general sets of classical probabilities are obtained by the theory of coherent previsions ([33]), i.e. by assigning interval-valued expectations $\mathbb{E}_{\mathcal{M}}(\cdot) := [{}^{\mathrm{L}}\mathbb{E}_{\mathcal{M}}(\cdot), {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\cdot)]$ on a set $\mathcal{K}$ of random variables on $(\Omega, \mathcal{A})$. By the lower envelope theorem ([33, p.134]) and the fact that classical expectation and classical probabilities uniquely correspond with each other, the definition of coherence can be rewritten in a way similar to (8). Since Walley [33] did not coin a name for the resulting set of classical probabilities, it will be called *structure*, too.

The interval-valued functions or functionals and the structure are dual concepts, they uniquely determine each other. The results obtained in this paper will be given in terms of the structure.

Many concepts of classical probability theory can be generalized appropriately. For decision making the notion of expectation is the most important one. Looking at the structure $\mathcal{M}$, one way how to define expectation for interval probability and how to extend the functional $\mathbb{E}_{\mathcal{M}}$ to random variables $X \notin \mathcal{K}$ suggests itself (see also the natural extension in [33]): Given a structure $\mathcal{M} \subseteq \mathcal{K}(\Omega, \mathcal{A})$

$$\mathbb{E}_{\mathcal{M}}X := \left[{}^{\mathrm{L}}\mathbb{E}_{\mathcal{M}}X, {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}X\right] := \left[\inf_{p(\cdot)\in\mathcal{M}} \mathbb{E}_p X\,,\ \sup_{p(\cdot)\in\mathcal{M}} \mathbb{E}_p X\right] \tag{9}$$

is the *(interval-valued) expectation of X (with respect to $\mathcal{F}$).*[4]

3.2 Generalized Expected Loss and Risk

In this section the decision problem as described in the Introduction will be analyzed in the situation where the decision maker's knowledge on the states of nature is ambiguous, expressed by a structure $\mathcal{M}$ of classical probabilities on $(\Theta, \mathcal{P}o(\Theta))$. To focus the argumentation on the essential ideas, it is assumed that the sampling information consists of classical probabilities.[5]

The generalization of the concept of probability now allows to consider generalized prior probabilities describing the decision maker's state of knowledge. With the notion of interval-valued expectation from (9) one immediately obtains the basic element of a generalized decision theory:

Definition 1 *Consider the basic decision problem $(\mathrm{IA}, \Theta, l(\cdot))$, a structure $\mathcal{M} \subseteq \mathcal{K}(\Theta, \mathcal{P}o(\Theta))$, and a sample information $(\mathcal{X}, \mathcal{P}o(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$. For every (randomized action) $\lambda \in \Lambda(\mathrm{IA})$, and every decision function $d \in \mathrm{ID}$, the expectations*

[4] An alternative way to define expectation for non-additive set functions is the *Choquet integral (or fuzzy integral)* (c.f., e.g., [9]). For the case of two-monotone and totally monotone capacities both notions are equivalent (cf., e.g., [9, Prop. 10.3, p. 126]). Therefore, the results developed below are then valid for the Choquet integral, too.

[5] The whole framework can be extended to imprecise sample information without substantial difficulties (cf., also the brief outline in [1]).

$\mathbb{E}_{\mathcal{M}}\mathbf{l}(\lambda)$ *and* $\mathbb{E}_{\mathcal{M}}\mathbf{R}(d)$ *are the* generalized expected loss *and the* generalized expected risk *(with respect to the prior information* $\mathcal{M}$*), respectively.*

Note that $\mathbb{E}_{\mathcal{M}}\mathbf{l}(\lambda)$ and $\mathbb{E}_{\mathcal{M}}\mathbf{R}(d)$ are interval-valued quantities. In most cases, comparing the generalized expected loss of actions directly will lead only to partial orderings on IA and $\Lambda(\mathrm{IA})$. If a linear (complete) ordering of actions is desired, an appropriate *representation* is needed. This is a mapping from $\mathbb{R}\times\mathbb{R}$ to $\mathbb{R}$ which evaluates intervals by real numbers to use the natural ordering on $\mathbb{R}$ for distinguishing optimal actions.

Expressing the probabilistic knowledge by a structure means that inside the structure there is complete ignorance: none of the elements of the structure is 'more likely' than another one. Therefore several authors (see the literature cited below) suggested to apply 'the maximin criterion to the structure'. Then the interval-valued expectations are represented by the upper interval limit alone. Accordingly, an action λ^* or a decision function d^* is optimal iff

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{l}(\lambda^*)) \leq {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{l}(\lambda))\,, \quad \forall\lambda\in\Lambda(\mathrm{IA})\,. \tag{10}$$

and

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{R}(d^*)) \leq {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{R}(d))\,, \quad \forall d\in\mathbb{D}\,, \tag{11}$$

respectively. The criterion (10) corresponds, among others, to the Maxmin expected utility model ([15]) and to the MaxEMin criterion considered by Kofler and Menges ([21]; cf. also [20] and the references therein)). (11) is also called Gamma-Minimax principle (e.g. [4, Section 4.7.6],[32]). These criteria will be used in this paper, too.[6]

Remark 2 *It should be noted that the criterion considered here contains the two main classical decision criteria as border cases: If there is perfect probabilistic information and therefore no ambiguity, then* $\mathcal{M}$ *consists of one single classical prior probability* $\pi(\cdot)$ *only; (10) and (11) coincide with Bayes optimality with respect to* $\pi(\cdot)$*. On the other hand, in the case of completely lacking information, the prior information consists of all classical probabilities on* $(\Theta,\mathcal{P}o(\Theta))$ *(*'non-selective' *or* 'vacuous' *prior). Then it is easily derived that*

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(l(\lambda)) = \min_{j\in\{1,\ldots,m\}} l(d,\vartheta_j) \quad \textit{and} \quad {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}}(\mathbf{R}(d)) = \max_{j\in\{1,\ldots,m\}} R(d,\vartheta_j)\,,$$

and (10) as well as (11) lead to the minimax criterion.

[6] This is done, however, without claiming that this is the only appropriate choice. Indeed, already in the seminal paper by Ellsberg [11] there are strong arguments for additionally taking into account other criteria. A convenient and nevertheless flexible choice is a linear combination of lower and upper limits (compare, e.g., with [11, p. 664], [18],[40], [39, Ch. 2.6]).

4 Robust Bayesian Analysis and Generalized Bayes Rule

4.1 Posterior Loss Analysis

The search for a decision function is much more costly than the calculation of optimal actions. Therefore, a natural attempt to solve (11) relies on the idea of the main theorem of Bayesian decision theory (compare Proposition 1): after having observed $\{x\}$, calculate the (now imprecise) posterior to update the imprecise prior, and then determine the action minimizing posterior loss.

Before discussing properties of this way to proceed in detail, the informal description just given has to be made precise:

Definition 2 *Consider the basic decision problem* $(\mathrm{IA}, \Theta, l(\cdot))$*, a structure* $\mathcal{M} \subseteq \mathcal{K}(\Theta, \mathcal{P}o(\Theta))$*, and a sample information* $(\mathcal{X}, \mathcal{P}o(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$*. Assume that* $\pi(\{\vartheta\}) > 0, \forall \vartheta \in \Theta,\ \forall \pi \in \mathcal{M}$.

i) Then, for every $x \in \mathcal{X}$*, call*

$$\mathcal{M}_{|x} = \{\pi(\cdot|x) | \pi \in \mathcal{M}\} \tag{12}$$

the imprecise posterior *given x, and* $\lambda^* \in \Lambda(\mathrm{IA})$ *with*

$$^{\mathrm{U}}\mathbb{E}_{\mathcal{M}_{\cdot|x}}(l(\lambda^*, \vartheta_j)) \leq {}^{\mathrm{U}}\mathbb{E}_{\mathcal{M}_{\cdot|x}}(l(\lambda, \vartheta_j)), \quad \forall \lambda \in \Lambda(\mathrm{IA}), \tag{13}$$

an optimal action with respect to the posterior loss *given x.*[7]

ii) A decision function $\tilde{d} = (\tilde{d}(x_1), \ldots, \tilde{d}(x_s))$ *where, for every* $s = 1, \ldots, k$*, the action* $\tilde{d}(x_s)$ *is optimal with respect to the posterior loss given* x_s*, is called* posterior loss optimal decision function*.*

The imprecise posterior from (12) is the main tool in robust Bayesian analysis (e.g., [35]), and its use is understood as self-evident in the decision theoretic work based on the theory of linear partial information ([21] and subsequent work). Moreover, a strong justification is provided by Walley's [33] theory. The calculation of $\mathcal{M}_{|x}$ is equivalent to applying his generalized Bayes rule, which is thoroughly derived from general axioms on coherent updating (cf. [33]). And indeed – next to its intuitive plausibility – working with the imprecise posterior has many further appealing properties. For instance, it is a vivid tool to reflect prior-data conflict ([33, p.6]) and it is naturally applied in successive updating where the imprecise posterior serves as an imprecise prior, once additional data are available.[8]

[7] Vidakovic [32] calls such optima *conditional* Gamma-Minimax solutions.

[8] See, however, [41, Section 6].

4.2 Suboptimality of Posterior Loss Optimal Decision Functions

Though this procedure seems to suggest itself, it must, however, be noted that its decision theoretic foundation is lost. As has to be discussed here, the decision function constructed along the lines of Part ii) of Definition 2 may be **suboptimal** with respect to the criterion (11).

A very simple counterexample can be obtained from a border case: Consider the vacuous prior information $\mathcal{K}(\Theta, \mathcal{P}o(\Theta))$. Then, independent of x, also the imprecise posterior is vacuous[9]. Using it as the 'updated prior' yields, for every x, according to Remark 2, the maximin solution λ^{mm} of the basic decision problem as the optimal randomized action. In contrast, the optimal decision function coincides with the maximin decision function $d^{mm}(\cdot)$ of the data problem. Typically, $d^{mm}(\cdot)$ has lower risk than the decision function $\tilde{d} = (\lambda^{mm}, \lambda^{mm}, \ldots, \lambda^{mm})$. Other counterexamples can be obtained, for instance, by considering situations, where the posterior probabilities are dilated (for this phenomenon see: [31, 36]).

The relation to minimax solutions goes far beyond the border case counterexample just given. Indeed, the following representation theorem even shows that optimal actions in the sense of (10) and optimal decision functions according to (11) *are minimax* solutions (in a different decision problem, where the structure serves as the set of states of nature) — except in the case of classical probability where the structure consists of a single element only. Therefore, the optimal solution must share all the (un)pleasant properties of minimax solutions, and so a reduction of the data problem to smaller basic decision problems cannot be expected; the equivalence of optimality with respect to posterior loss and to prior risk has to be given up.[10]

Theorem 1 (Representation Theorem) *Consider the basic decision problem* $(\mathrm{IA}, \Theta, l(\cdot))$*, the prior structure* $\mathcal{M} \subseteq \mathcal{K}(\Theta, \mathcal{P}o(\Theta))$*, and a sample information* $(\mathcal{X}, \mathcal{P}o(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$*. Then the following equivalences hold:*

i) An action λ^* *is optimal with respect to the criterion (10), iff it is minimax action in the basic decision problem* $(\Lambda(\mathrm{IA}), \mathcal{M}, \tilde{l}(\cdot))$ *with*

$$\begin{array}{rcccl} \tilde{l} & : & (\Lambda(\mathrm{IA}) \times \mathcal{M}) & \to & \mathbb{R} \\ & & (\lambda, \pi) & \mapsto & \tilde{l}(\lambda, \pi) := \mathbb{E}_\pi(l(\lambda, \vartheta)) . \end{array}$$

ii) A decision function $d^*(\cdot)$ *is optimal with respect to the criterion (11), iff* $d^*(\cdot)$ *is minimax solution in the basic decision problem* $(\mathcal{D}, \mathcal{M}, \tilde{R}(\cdot))$ *with*

$$\begin{array}{rcccl} \tilde{R} & : & (\mathcal{D} \times \mathcal{M}) & \to & \mathbb{R} \\ & & (d, \pi) & \mapsto & \tilde{R}(d, \pi) := \mathbb{E}_\pi(R(d, \vartheta)) . \end{array}$$

[9] See, for instance, [33, p.308].

[10] For the same reason also the essential completeness of unrandomized actions, known from classical Bayesian theory, is no longer valid.

Sketch of the proof: For Part i) read the criterion (10)

$$\max_{\pi(\cdot)\in\mathcal{M}} \mathbb{E}_\pi(l(\lambda,\vartheta)) \to \min$$

from the viewpoint of the minimax criterion (4), where Θ has been replaced by $\mathcal{M}$. To show Part ii), analogously rewrite (11) in the light of (5).

The basic idea of this theorem is similar to Schneeweiß' [27] representation of a basic decision problem. A closer study of the proof shows that this theorem can also be directly extended to imprecise sample information and to the Hurwicz-like optimality criteria briefly mentioned in Footnote 6. Moreover, the fact that in this representation the structure $\mathcal{M}$ now serves as the set of states of nature provides straightforwardly a framework for decision making with second order probabilities: in this setting, a prior weighing the states of nature is nothing but a second order distribution.

5 Concluding Remarks

The paper showed that, for imprecise probability, optimality with respect to prior risk and to posterior loss need no longer coincide. Decision functions constructed by collecting, for every potential observation $x \in \mathcal{X}$, the optimal actions given the corresponding imprecise posterior structure may have higher risk than the direct solution to (11). From the computational point of view this means that, in order to calculate the risk minimizing solution, the reduction to small, easy to solve basic decision problems, which is characteristic for the Bayesian approach in the classical setting, is not possible any more; it is indispensable to go the costly way, fraught with difficulty, via the optimal decision *function*. Efficient algorithms solving this challenge in contexts of optimal design and testing are provided by Fandom Noubiap and Seidel [12, 13]. Augustin [1, 3] gives a general algorithm which is, in principle, applicable to arbitrary decision problems on finite spaces.

Concerning the foundations of statistics it is remarkable that, in the area of imprecise probabilities, the intensive debate between frequentists and Bayesians on topics like counterfactual effects and the principle of conditionality, obtains new importance. Should inference be based only on the concrete observation x, or should one take all potential observations $x \in \mathcal{X}$ into account, i.e., evaluate the decision function as a whole? There are sound arguments for both views and, quite evidently, the author is not the one to decide the question definitely. But, at least, it can be said that one should be aware of the fact that in the area of imprecise probability, in contrast to classical theory, now the standpoint matters; it may influence the results substantially. The imprecise posterior does no longer contain all the relevant information to produce optimal decisions. Inference and decision do not coincide any more — just as in every day life, there is a difference between accumulating as much information as possible (inference and updating knowledge) and making optimal decisions. This may lead to a number of paradoxes, since

statisticians up to now have been used to phrase estimating and testing problems equivalently as inference as well as decision problems.

Important further insights into the topic should arise from a deeper understanding of the relationship between the result obtained here and the phenomenon of dilation in conditioning imprecise probabilities as described by Seidenfeld and Wasserman [31] and Wasserman and Seidenfeld [36]. There should also be a close and illuminating connection to Jaffray's [19] observations on sequential decision making, and to Seidenfeld's paper ([29]) on incoherence in sequential decision making when preferences fail the independence axiom.[11]

Further research may also attempt at reconciling the conditional and the so-to-say global point of view, the more as the debate on appropriately defining conditional imprecise probabilities is far from being closed. An increasing number of results supports the idea that there should be a symbiosis of several concepts of conditional interval probability ([10, 16, 41] and the references provided there.). There may be some hope to find a notion of conditional probability or a meaningful optimality criterion under which both ways to proceed coincide. In such a setting there would be unanimity on the meaning of terms like 'updating', 'inference' and 'optimal decision making', because then, and only then, the posterior would contain all the relevant information for decision making.

Acknowledgement

I am very grateful to numerous colleagues for clarifying discussions on this topic, which has haunted me for quite a long time. Special thanks go to Kurt Weichselberger, Franz Ferschl, Toni Wallner, Bernhard Rüger, Frank Coolen, Gert de Cooman and Teddy Seidenfeld. My appreciation extends to an anonymous member of the program board and to three anonymous referees.

References

[1] Augustin, T. (2001): On decision making under ambiguous prior and sampling information. In [6], 9-16.

[2] Augustin, T. (2002): Expected utility within a generalized concept of probability — a comprehensive framework for decision making under ambiguity. *Statistical Papers* 43, 5-22.

[3] Augustin, T. (2003): Optimal decisions under complex uncertainty — basic notions and a general algorithm for data-based decision making with partial prior knowledge described by interval probability. Submitted. `www.stat.uni-muenchen.de/~thomas/augustin2003d.pdf`

[11] Concerning the relation between the independence axiom and dilation see [30, Section 3.1].

[4] Berger, J.O. (1984): *Statistical Decision Theory and Bayesian Analysis.* (2nd edition). Springer. New York.

[5] de Cooman, G.,Cozman, F.G., Moral, S. and Walley, P. (Eds.) (1999): *ISIPTA'99: Proceedings of the First International Symposium on Imprecise Probabilities and their Applications. Ghent.*

[6] de Cooman, G., Fine, T.L., Moral, S., and Seidenfeld, T. (Eds.) (2001): *ISIPTA 01: Proceedings of the Second International Symposium on Imprecise Probabilities and their Applications. Cornell University, Ithaca (N.Y.)*, Shaker, Maastricht.

[7] de Cooman, G., and Walley, P. (Eds.) (2003): *The Imprecise Probability Project.* `http://www.sipta.org`.

[8] Cozman, F.G. (2000): Computing posterior upper expectations. *Internatiopnal Journal of Approximate Reasoning* 24, 191-205.

[9] Denneberg, D. (1994): *Non-Additive Measure and Integral.* Kluwer. Dordrecht.

[10] Dubois, D., and Prade, H. (1994): Focusing versus updating in belief function theory. In: R.R. Yager, M. Fedrizzi and J. Kacprzyk (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*. Wiley, New York, 71-95.

[11] Ellsberg, D. (1961): Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics* **75**, 643-669.

[12] Fandom Noubiap, R., and Seidel, W. (2001a): An algorithm for calculating Γ-minimax decision rules under generalized moment conditions. *Annals of Statistics* 29, 1094-1116.

[13] Fandom Noubiap, R., and Seidel, W. (2001b): An efficient algorithm for constructing Γ-minimax tests for finite parameter spaces. *Computational Statistics and Data Analysis* 36, 145-161.

[14] Ghirardato, P. and Marinacci, M. (2001): Risk, ambiguity, and the separation of utility and beliefs. *Mathematics of Operations Research* 26, 864-890.

[15] Gilboa, I., and Schmeidler, D. (1989): Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* 18, 141–153.

[16] Halpern, J.Y., and Fagin, R. (1992): Two views of belief: belief as generalized probability and belief as evidence. *Artificial Intelligence* 54, 275-317.

[17] Huber, P.J., and Strassen, V. (1973): Minimax tests and the Neyman-Pearson lemma for capacities. *Annals of Statistics* 1, 251-263; Correct.: 2, 223-224.

[18] Jaffray, J.Y. (1989): Linear utility theory and belief functions. *Operations Research Letters* 8, 107–122.

[19] Jaffray, J.Y. (1999): Rational decision making with imprecise probabilities. In [5], 183-188.

[20] Kofler, E. (1989): *Prognosen und Stabilität bei unvollständiger Information*. Campus. Frankfurt/Main.

[21] Kofler, E., and Menges, G. (1976): *Entscheidungen bei unvollständiger Information.* Springer. Berlin.

[22] Kuznetsov, V.P. (1991): *Interval Statistical Methods.* Radio i Svyaz Publ., (in Russian).

[23] Levi, I. (1974): On indeterminate probabilities. *Journal of Philosophy* 71, 391-418.

[24] Levi, I. (1980): The Enterprise of Knowledge. An Essay on Knowledge, Credal Probability and Chance. MIT Press, Cambridge (MA).

[25] Levi, I. (1990): Consequentialism and sequential choice. In: Bacharch, M. and Hurley, S. (Eds.): *Foundations of Decision Theory.* Blackwells, Oxford; 92-122.

[26] Rios Insua, D.R., and Ruggeri, F. (Eds.) (2000): *Robust Bayesian Analysis.* Springer (Lecture Notes in Statistics 152), New York.

[27] Schneeweiß, H. (1964): Eine Entscheidungsregel im Fall partiell bekannter Wahrscheinlichkeiten. *Unternehmensforschung* 10, 86-95.

[28] Shafer, G. (1976): *A Mathematical Theory of Evidence.* Princeton University Press. Princeton.

[29] Seidenfeld, T. (1988): Decision theory without 'independence' or without 'ordering'. Waht is the difference?. *Economics and Philosophy* 4, 267-290.

[30] Seidenfeld, T. (1994): When normal and extensive form decisions differ. In: Prawitz, D., Skyrms, B. and Westerstahl, D. (Eds.): *Logic, Methodology and Philosophy of Science IX (Uppsala, 1991).* Elsevier, Amsterdam, 451-463

[31] Seidenfeld, T., and Wasserman, L. (1993): Dilation for sets of probabilities. *Annals of Statistics* 21, 1139-1154.

[32] Vidakovic, B. (2000): Γ-minimax: A paradigm for conservative robust Bayesians. In: Insua, D.R., and Ruggeri, F. (Eds.): *Robust Bayesian Analysis.* Springer (Lecture Notes in Statistics 152), New York, 241-259.

[33] Walley, P. (1991): *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall. London.

[34] Walley, P., and Fine, T.L. (1982): Towards a frequentist theory of upper and lower probability. *The Annals of Statistics* 10, 741-761.

[35] Wasserman, L. (1997): Bayesian robustness. In: S. Kotz, C.B. Read, D.L. Banks (Eds.): *Encyclopedia of Statistical Sciences. Update Volume 1.* Wiley, New York, pp. 45-51.

[36] Wasserman, L., and Seidenfeld, T. (1994): The dilation phenomenon in robust Bayesian inference. (With discussion). *Journal of Statistical Planning and Inference* 40, 345-356.

[37] Weichselberger, K. (1995): Axiomatic foundations of the theory of interval-probability. In: V. Mammitzsch, and H. Schneeweiß (Eds.): *Symposia Gaussiana Conference B.* de Gruyter, Berlin, 47-64.

[38] Weichselberger, K. (2000): The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning* 24, 149-170.

[39] Weichselberger, K. (2001): *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept.* Physika, Heidelberg.

[40] Weichselberger, K., and Augustin, T. (1998): Analysing Ellsberg's Paradox by means of interval-probability. In: R. Galata, and H. Küchenhoff (Eds.): *Econometrics in Theory and Practice. (Festschrift for Hans Schneeweiß.)* Physika. Heidelberg, 291–304.

[41] Weichselberger, K., and Augustin, T. (2003): On the symbiosis of two concepts of conditional interval probability. Conditionally accepted for: Bernard, J.M., Seidenfeld, T., and Zaffalon, M. (Eds.): *ISIPTA 03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications, Lugano.*

[42] Yager, R.R., Fedrizzi, M., and Kacprzyk, J. (Eds.) (1994): *Advances in the Dempster-Shafer Theory of Evidence.* Wiley. New York.

T. Augustin is with the Department of Statistics, Ludwig-Maximilians University of Munich, Ludwigstr. 33, D-80539 München, Germany. E-mail: augustin@stat.uni-muenchen.de