J. R. Statist. Soc. B (1993)
55, *No.* 1, *pp.* 253–266

On the Empirical Validity of the Bayesian Method

By V. G. VOVK† and V. V. V'YUGIN

Research Council for Cybernetics, Moscow, USSR

[Received January 1990. Final revision October 1991]

SUMMARY

The ideas in Kolmogorov's programme for algorithmic substantiation of applications of probability make it possible to define a measure of disagreement between the probability distribution representing the attitude of a coherent individual towards a random experiment and the outcome of the experiment. When there is agreement we say that the probability distribution is empirically valid. We prove quantitatively that formulae of Bayesian statistics transform empirically valid probability distributions into other empirically valid distributions.

Keywords: ALGORITHMS; BAYESIAN STATISTICS; FALSIFICATION; KOLMOGOROV'S PROGRAMME

1. INTRODUCTION

Let Ω be the set of all possible outcomes of some random experiment (endowed with a σ-algebra). de Finetti (1937), section I, has shown that a coherent individual's judgments concerning this experiment can be described by a probability distribution P in Ω. (Strictly, P is guaranteed to be only finitely additive; however, we shall usually consider situations where finite additivity implies σ-additivity.) We know that people sometimes admit that their *a priori* beliefs have been wrong. For example, the holder of P might find the real outcome ω of a random experiment to be morally impossible (under P) and thus reject P as empirically invalid. It seems that we need some criterion of disagreement between P and ω.

There are various attitudes towards disagreement. de Finetti believes that 'observation [i.e. ω] cannot confirm or refute an opinion [i.e. P]' (de Finetti (1937), section VI). The opposite attitude has been expressed by Dawid: 'We are thus led to consider any subjective model, or parametric model, of the behaviour of observables as a meaningful theory in the sense of Popper: we can conceive of experimental data that would discredit it' (Dawid (1986), section 6).

We start with the assumption that some criterion of disagreement between P and ω is needed. There are two sides to the disagreement between P and ω: we can say either that ω falsifies P or that ω is morally impossible under P.

In statistics definite falsification is usually unattainable, so rather we need a quantitative measure of disagreement between P and ω. Unfortunately, our definition (which is a slight variation of known definitions—see later) of the measure of disagreement relies heavily on the theory of algorithms, which is unfamiliar to most statisticians. So we shall explain the main ideas in a situation that is as simple as possible. In particular, we shall confine ourselves to the case of a single individual.

†*Address for correspondence*: 9-3-451 ulitsa Ramenki, Moscow 117607, USSR.

 0035–9246/93/55253 $2.00

Let P be a probability distribution in Ω describing beliefs of some coherent individual. In Section 3.1 we define a function $p(\omega)$, $\omega \in \Omega$, which, as we show, can be interpreted as an 'essentially best' measure of impossibility of ω under P (we prefer to speak of 'impossibility' instead of 'disagreement' since P is fixed and ω is the only argument of p). Outcomes ω with large $p(\omega)$ are interpreted as almost impossible from the viewpoint of the holder of P. We call p the level of impossibility. Our definition presupposes that the distribution P is computable in a reasonable way (according to Dawid (1985a), 'it is reasonable to claim that any possible statistical analysis, formal or informal, must be computable').

In Section 3.2 analogous definitions are given in a situation that is usual in Bayesian statistics. Let $Q(\mathrm{d}\theta)$ be the probability distribution in a parameter set which characterizes beliefs of the individual. For each parameter value θ, let P_θ be the probability distribution in Ω describing his beliefs concerning the random experiment with outcomes in Ω conditional on knowing θ (we suppose that it makes sense to speak of the 'true value' of the parameter).

The distributions Q and (P_θ) determine the joint distribution $T(\mathrm{d}\theta, \mathrm{d}\omega)$ of θ and ω, the Bayesian mixture $Y(\mathrm{d}\omega)$ and (under conditions of regularity) the posterior distributions $\{Z_\omega(\mathrm{d}\theta)\colon \omega \in \Omega\}$. Let A and B range over measurable sets of parameter values and outcomes respectively. The distribution T is defined by the equality

$$T(A \times B) = \int_A P_\theta(B)\, Q(\mathrm{d}\theta),$$

Y by $Y(B) = T(\Theta \times B)$, where Θ is the parameter set, and (Z_ω) by the requirement that

$$T(A \times B) = \int_B Z_\omega(A)\, Y(\mathrm{d}\omega).$$

The interpretation of T, Y and Z_ω is as follows: $T(A \times B)$ is the degree of the individual's belief that simultaneously $\theta \in A$ and $\omega \in B$, $Y(B)$ is the degree of his belief that $\omega \in B$ and $Z_\omega(A)$ is the degree of his belief that $\theta \in A$ after the outcome ω is known.

Let θ be the true value of the parameter and ω the outcome of the experiment. Suppose that θ and ω agree with Q and P_θ respectively. Then, as is proved in Section 4, the pair (θ, ω) agree with T, ω agrees with Y and θ agrees with Z_ω. In other words, the transition from P, Q to T, Y, Z conserves the empirical validity. The quantitative form of this assertion has interesting consequences.

Let us briefly review some relevant publications. We distinguish two principal trends in the development of ideas connected with the level of impossibility; we call them the stochastic approach (following Kolmogorov and Uspenskii (1987)) and the measure theoretic approach. The former was originated by R. von Mises and the latter by A. N. Kolmogorov. In this paper we are concerned only with the measure theoretic approach (the stochastic approach in its advanced form lies outside the scope of Bayesian statistics).

The measure theoretic approach has been put forward by Kolmogorov (1965), section 4, and Martin-Löf (1966). Kolmogorov proposes the following definition: an element of a finite set is random (in our nomenclature, agrees with the uniform probability distribution in this set) if its algorithmic complexity (defined in Kolmogorov (1965)) is close to the largest possible value. The difference between this largest possible value and the algorithmic complexity can serve as a measure of impossibility of the element. Kolmogorov and Uspenskii (1987) called this difference the deficiency of randomness of the element.

Martin-Löf (1966), section 2, noted that Kolmogorov's definition can be usefully reformulated in terms of a universal test for randomness. The definition so reformulated was easy to transfer to the case of the space Ω of all infinite 0–1-sequences. In section 3 of this work Martin-Löf gives the definition of a measure of impossibility with respect to the uniform probability distribution in Ω, and in section 4 this definition is extended to arbitrary computable probability distributions in Ω (even in the latter case Martin-Löf continues to use the term 'randomness').

Kolmogorov's original idea of basing the notion of randomness on that of algorithmic complexity has been also developed (see Kolmogorov and Uspenskii (1987), section 1.4, and the references therein). Despite all the intuitive appeal of this approach, its technically feasible generalizations tend to lead to either a measure theoretic (Martin-Löf, 1966) or a stochastic (see Kolmogorov and Uspenskii (1987), section 1.4, remark 2) approach. However, Kolmogorov's idea is implicit in Rissanen's influential 'coding' approach to statistical inference (see, for example, Rissanen (1987)).

2. ALGORITHMIC BACKGROUND

Any satisfactory definition of level of impossibility seems bound to involve elements of the theory of algorithms. This theory is systematically treated in, for example, Rogers (1967). However, to understand this paper an intuitive idea of algorithms will suffice.

In this paper we need only algorithms which transform a finite object and an infinite 0–1-sequence into a finite object. Integer and rational (but not real) numbers are examples of finite objects. Finite sequences of finite objects are again finite objects. Infinite 0–1-sequences are taken in by algorithms sequentially bit by bit. In particular, the output of an algorithm fed with a sequence ω is determined by some finite prefix of ω (i.e. the output will be the same when the algorithm is fed with any other continuation of this prefix). An (admissible) *input* is a pair consisting of a finite object (the *finite* part of the input) and an infinite 0–1-sequence (the *infinite* part of the input).

Let $\overline{\mathbb{R}}$ denote the real line extended by adding the infinities $-\infty$ and ∞. A function $f: A \rightarrow \overline{\mathbb{R}}$, where A is a set of inputs, is *upper* or *lower semicomputable* if there is an algorithm $\mathscr{U}$ which, when fed with a rational number r and an input $a \in A$ (the pair (r, a) can be considered as a new input), eventually stops if $f(a) < r$ or $f(a) > r$ respectively and never stops otherwise. (Sometimes we shall say that f is upper or lower semicomputable by $\mathscr{U}$.) In other words, the lower semicomputability of f means that if $f(a) > r$ this fact will sooner or later be learned (it is *positively decidable*), whereas if $f(a) \leqslant r$ we may be for ever uncertain (this inequality may not be positively decidable), and analogously for upper semicomputability. The function f is *computable* if some algorithm transforms any input $a \in A$ and positive integer n into a rational number r satisfying $|f(a)-r| \leqslant 2^{-n}$, i.e. f is computable if the value $f(a)$ can be computed arbitrarily accurately. If A is a set of finite objects or infinite 0–1-sequences, the definition is the same except that the algorithm is required to ignore the infinite or finite part of the input respectively. We also allow situations where A consists of, for example, pairs of infinite 0–1-sequences: such a pair $a_1a_2\ldots$, $b_1b_2\ldots$ can be represented as the single 0–1-sequence $a_1b_1a_2b_2\ldots$.

Lemma 1. A function $f: A \to \bar{\mathbb{R}}$, A being a set of inputs, is computable if and only if it is simultaneously upper semicomputable and lower semicomputable.

The next lemma, which asserts the existence of a *universal* lower semicomputable function f, is of fundamental importance. Let $\mathbb{N}$ denote the positive integers.

Lemma 2. For any set A of inputs there is a lower semicomputable function f: $\mathbb{N}\times A \to \bar{\mathbb{R}}$ such that the sequence $f_1, f_2, \ldots$ of the functions $f_n(a) = f(n, a)$ contains all lower semicomputable functions of the type $A \to \bar{\mathbb{R}}$.

Proof. Each algorithm is described by a *program* which is a finite sequence of symbols in some alphabet. (In practice not all sequences are 'meaningful' programs, but it is convenient to consider 'meaningless' programs as programs describing algorithms that never stop.) Let $\pi_1, \pi_2, \ldots$ be a computable enumeration of all programs. We define an algorithm $\mathscr{U}$ (for checking whether $f(n, a) > r$ holds) as follows. When fed with n, a and r, it applies the program π_n to all pairs (a, r') such that $r' > r$ and r' is a rational number; $\mathscr{U}$ stops when at one of these pairs the program π_n stops. It is easy to check that there is a (unique) function $f(n, a)$ that is lower semicomputable by $\mathscr{U}$. Any lower semicomputable function g: $A \to \bar{\mathbb{R}}$ is computed by some program π_n and, therefore, g coincide with f_n. □

Let Ω denote the set of all infinite 0–1-sequences. If x is a finite 0–1-sequence, we define $\Gamma(x) \subseteq \Omega$ as the set of all $\omega \in \Omega$ such that x is a prefix of ω. A function f on Ω taking rational or infinite ($-\infty$ or ∞) values is *simple* if there is a partition of the set Ω into sets $\Gamma(x_1), \ldots, \Gamma(x_n)$ such that f is constant on each of the $\Gamma(x_i)$. Note that simple functions are finite objects. A sequence $f_1, f_2, \ldots$ of finite objects is computable if some algorithm transforms arbitrary $n = 1, 2, \ldots$ into f_n. The next lemma gives a useful characterization of lower semicomputable functions.

Lemma 3. For any lower semicomputable function f on Ω there is a computable non-decreasing sequence of simple functions f_n such that $f_n(\omega)$ tends to $f(\omega)$ (as $n \to \infty$) for all ω.

Proof. Let f be lower semicomputable by an algorithm $\mathscr{U}$. The algorithm for computing f_n given n acts as follows. For each pair (x, r), where x is a finite 0–1-sequence and r is a rational number, it models the work of $\mathscr{U}$ on the sequence $x00\ldots$ and the number r. When it sees that $\mathscr{U}$ stops without taking in 0s outside the part x of the sequence $x00\ldots$, it stores the pair (x, r). (All this should be done independently of n.) After n steps of work it outputs the following function f_n: for any $\omega \in \Omega$, $f_n(\omega)$ is defined as $\sup(r_i)$, where r_i are the second elements of the stored pairs (x, r) in which x is a prefix of ω ($\sup(\varnothing)$ is defined to be $-\infty$). □

Lemma 3 can be easily generalized to arbitrary lower semicomputable functions. For simplicity we prove lemma 1 only in the case $A = \Omega$.

Proof of lemma 1 for $A = \Omega$. We only need to prove that any function f which is both upper and lower semicomputable is computable. By lemma 3 there are computable non-decreasing sequences f_n and $-g_n$ of simple functions which converge to f and $-f$ respectively (so g_n is a non-increasing sequence which converges to f). To compute a rational number r satisfying $|f(\omega)-r| \leqslant 2^{-m}$ it suffices to find n satisfying $|f_n(\omega)-g_n(\omega)| \leqslant 2^{-m}$ and output $f_n(\omega)$. □

3. LEVEL OF IMPOSSIBILITY

3.1. *Part I*

The aim of this section is to define the level of impossibility p with respect to a computable probability distribution P in the set Ω of infinite 0–1-sequences. (The computability of P means that the function $P\{\Gamma(x)\}$, x ranging over finite 0–1-sequences, is computable.) We often interpret Ω as the closed interval [0, 1] associating with each 0–1-sequence $\omega_1\omega_2 \ldots$ the real number with the binary expansion $0.\omega_1\omega_2 \ldots$. (Note, however, that the numbers in the open interval]0, 1[with a finite binary expansion, which are in a certain sense few, are represented in Ω by two sequences; for example, 0.1 is represented by 100. . . and 011. . . .)

Recall that p measures the empirical invalidity of P. In accordance with Popper (1934), chapter VI, the empirical character of the level of impossibility means that it should determine potential falsifiers $\omega \in \Omega$ of P. So the interpretation of p will be as follows: P is falsified by an outcome ω (or ω is impossible with respect to P) at a level α if $p(\omega) > \alpha$ (by Popper (1934), chapter VIII, probability statements are impervious to strict falsification, and we must content ourselves with falsification at some level). The fact that P is falsified by ω at a level α should be positively decidable (if it is impossible to establish falsification effectively, then P is not falsified yet). So we have the following requirement: the level of impossibility should be lower semicomputable. (Note the asymmetry between lower and upper semicomputability here, which is a consequence of the asymmetry between falsification and verification.)

The requirement of lower semicomputability of p is general and does not depend on the probability distribution P. How can we interpret the assertion that outcomes ω with large values $p(\omega)$ are hardly possible from the point of view of the holder of P? We shall suppose that the holder of P is infinitely rich and immortal (so that utilitarian considerations will not prevent him from disclosing his true beliefs).

Let us borrow ideas from physics where one studies an electric field by placing into it a unit electric charge. Suppose that we have £1 (an analogue of the charge). If f is a non-negative function such that $\int f\,\mathrm{d}P < 1$, the holder of P should accept the following proposal: we pay him our £1 before the outcome of the experiment is known, and he pays us £$f(\omega)$ as soon as the outcome ω is known (f may be unlimited: recall that the holder of P is infinitely rich). Our next requirement is that $\int p\,\mathrm{d}P < 1$. The readiness of the holder of P to pay us a big amount $p(\omega)$ when ω occurs can be interpreted as the moral impossibility of ω from his point of view. Lower semicomputable non-negative functions f satisfying $\int f\,\mathrm{d}P < 1$ will be called *measures of impossibility* (with respect to P). Summarizing, we require that p be a measure of impossibility.

Remark 1. There arises a natural question whether a measure of impossibility f, which may be non-computable, can be really interpreted as a possible pay-off of the individual. We propose the following scenario. We give the individual £1 and a computer program (see lemma 3) for computing a non-decreasing sequence $f_1, f_2, \ldots$ of non-negative simple functions which converge to f, on the following condition. After producing each f_n we together with the individual check whether $\int f_n\,\mathrm{d}P < 1$. If yes, he pays us

$$£\{f_n(\omega)-f_{n-1}(\omega)\}$$

(where $f_0=0$) to bring the overall amount that he has paid us to $£f_n(\omega)$. If no, the agreement is annulled. (When $\int f_n \,\mathrm{d}P$ is exactly 1, there is a possibility that we shall never be able to decide which of the inequalities 'less than' or 'greater than or equal to' holds true. However, this causes no difficulties.)

There remains the question which of the measures of impossibility to choose as the level of impossibility. Following Popper (1934), chapter VI, we prefer a measure f_1 to a measure f_2 if f_1 'prohibits' more than f_2 does, i.e. we prefer f_1 to f_2 if

$$f_2(\omega) > \alpha \Rightarrow f_1(\omega) > \alpha, \qquad \text{for all } \omega, \alpha,$$

or, equivalent, if $f_1 \geqslant f_2$.

The next lemma, which is a variant of, for example, Schnorr (1977), proposition 3.8, asserts that there is an almost best measure of impossibility. Let $A = A(x_1, \ldots, x_n)$ and $B = B(x_1, \ldots, x_n)$ be non-negative numerical expressions with parameters $x_1, \ldots, x_n$. The inequality $A \preccurlyeq B$ (or $B \succcurlyeq A$) means that there is a constant $c > 0$ such that $A \leqslant cB$ for all $x_1, \ldots, x_n$; $A \asymp B$ means that $A \preccurlyeq B$ and $B \preccurlyeq A$.

Lemma 4. There is a (universal) measure of impossibility p such that $p \succcurlyeq p'$ for any other measure of impossibility p'.

Proof. Let $f\colon \mathbb{N} \times \Omega \to \overline{\mathbb{R}}$ be a universal lower semicomputable function and $f_n(\omega) = f(n, \omega)$ for all n, ω (see lemma 2). By lemma 3 for each n there is a non-decreasing computable sequence $g_{n1}, g_{n2}, \ldots$ of non-negative simple functions which converge to f_n^+ (defined by $f_n^+(\omega) = \max\{f_n(\omega), 0\}$). The proof of lemma 3 shows that g_{ni} can be assumed to be a computable function of the two arguments n and i. Put $g_{n\infty} = f_n^+$. For each n define

$$i(n) = \sup\left(i\colon \int g_{ni} \,\mathrm{d}P < 1\right).$$

Now it suffices to put $p = \Sigma\, 2^{-n-1} g_{n,i(n)}$. If p' is any other measure of impossibility, then $p' = f_n$ for some n and we have

$$p' = f_n^+ = g_{n\infty} = g_{n,i(n)} \leqslant 2^{n+1} p. \qquad \square$$

We fix a universal measure of impossibility p and call it the *level of impossibility* (with respect to P). Note that $p(\omega) \succcurlyeq 1$ (since the constant $\frac{1}{2}$ is a measure of impossibility). This definition is essentially due to Levin (1984). The requirements that p be lower semicomputable and that $\int p \,\mathrm{d}P < 1$ occur in Gacs (1980). It can be proved that the sequences ω satisfying $p(\omega) < \infty$ are exactly those random with respect to P in the sense of Martin-Löf (1966).

Remark 2. Chebyshev's inequality shows that, for any $\alpha > 0$,

$$P\{\omega\colon p(\omega) > \alpha\} < 1/\alpha.$$

So the probability that the level of impossibility of the outcome will be large is small. In particular, $p(\omega) < \infty$, P almost surely. Many results of probability theory are of the form '$A(\omega)$ holds P almost surely', where A is some property (e.g. the equality from the law of the iterated logarithm). One of the ways of strengthening such a result is to prove its 'pointwise' counterpart '$A(\omega)$ holds for *all* ω with $p(\omega) < \infty$' (see, for example, Kolmogorov and Uspenskii (1987) and Dawid (1985b)). The quantitative measure of impossibility p makes it possible to consider pointwise analogues of the assertions of the type 'the P probability of violating $A(\omega)$ is small' (such assertions are common in statistics): we can try to prove that $A(\omega)$ holds for *all* ω with small $p(\omega)$ (an example of such a pointwise assertion is given in remark 6 later).

Remark 3. Define

$$p^*(\omega) = \sup\{2^k : p(\omega) > 2^k, k \text{ is an integer}\}.$$

The function $p^*(\omega)$ is lower semicomputable as well as $p(\omega)$ and is connected with $p(\omega)$ by

$$\tfrac{1}{2}\, p(\omega) \leqslant p^*(\omega) \leqslant p(\omega).$$

Thus $p^*(\omega)$ is also a universal measure of impossibility. So without loss of generality we can assume that the level of impossibility takes values only of the form 2^k with integer k.

Popper's (1934) observation that verification is usually impossible is also valid in our case. An upper semicomputable function $f: \Omega \rightarrow \bar{\mathbb{R}}$ is called a *verifier* if $f(\omega) \geqslant p(\omega)$, for all ω. Let f be upper semicomputable by an algorithm $\mathscr{U}$. The $\mathscr{U}$ can be used for establishing agreement between P and an outcome ω: if $\mathscr{U}$ stops when fed with ω and a rational number r, then ω and P agree at level r (i.e. $p(\omega) < r$). There are no non-trivial verifiers.

Lemma 5. Every verifier identically equals ∞.

Proof. Let $f(\omega) < \infty$ for some ω. Then there exist a finite 0–1-sequence x and a constant c such that $f(\omega) < c$ for all $\omega \in \Gamma(x)$. This contradicts the evident fact that some measure of impossibility is unbounded on $\Gamma(x)$. □

So far we have ignored the fact that the level of impossibility depends on *a priori* information (this notion has been formally considered by Dawid (1985b), section 9). (To put it differently, we have supposed that no *a priori* information is available.)

Example 1. Let P be the uniform probability distribution in Ω. Consider a random experiment conducted according to P. If the individual is told of a sequence $\omega \in \Omega$ before the experiment, he will be sure that the outcome of the experiment will be different from ω, even if $p(\omega)$ is small.

We confine ourselves to the case that *a priori* information ι can be represented as a finite object. Considerations parallel to those at the beginning of this section lead to the following definition (the adjective 'conditional', which is used to distinguish the notions that we are defining now from the previous notions, has nothing in common with conditional distributions). A *conditional measure of impossibility* (with respect to P) is a lower semicomputable non-negative function $f(\omega \mid \iota)$, where ω and ι range over infinite 0–1-sequences and finite objects respectively, satisfying the inequality

$$\int f(\omega \mid \iota)\, P(\mathrm{d}\omega) < 1$$

for all ι. There are universal (i.e. largest to within a constant factor) conditional measures of impossibility, one of which we fix, call the *conditional level of impossibility* and denote by p. The value $p(\omega \mid \iota)$ is called the *level of impossibility* of ω relative to *a priori* information ι.

For example 1, recall that P is the uniform probability distribution in Ω. We have $p(\omega \mid \omega^n) \geqslant 2^n$, where $\omega \in \Omega$, $n \in \mathbb{N}$ and ω^n is the prefix of ω of length n (indeed, the function $f(\omega \mid \iota)$ equal to 2^{n-1}, if ι is a 0–1-sequence of length n and ω is a continuation of ι, and equal to 0 otherwise, is a conditional measure of impossibility). So the individual, when given a long 0–1-sequence before the experiment, is morally certain that this sequence will not be realized.

The next assertion says that scarce *a priori* information cannot considerably change the level of impossibility.

Lemma 6. When k ranges over positive integers,

$$p(\omega) \preccurlyeq p(\omega \mid k) \preccurlyeq k^2 p(\omega).$$

Proof. The left-hand inequality follows from the fact that the function $f(\omega \mid \iota)$ defined as $p(\omega)$ is a conditional measure of impossibility. Let us prove the right-hand inequality. The function $f(\omega)$ defined as $\Sigma\,(k+1)^{-2} p(\omega \mid k)$ is a measure of impossibility. Thus $p(\omega) \succcurlyeq f(\omega)$, whence $(k+1)^2 p(\omega) \succcurlyeq p(\omega \mid k)$. $\square$

Remark 4. k^2 in the statement of lemma 6 can be replaced by ν_k, where ν is any computable sequence such that the series $\Sigma\,\nu_k^{-1}$ converges. In particular, we can take

$$\nu_k = k \log_* k \log_*^{(2)} k \ldots \log_*^{(c-1)} k \,\{\log_*^{(c)} k\}^2,$$

where c is any fixed positive integer, $\log_* r$ is defined as 1 if $r \leqslant 2$ and as $\log r$ otherwise (the last logarithm is to the base 2) and $\log_*^{(i)}$ stands for $\log_* \ldots \log_*$ (i times).

Remark 5. Let P be the uniform probability distribution in Ω and ω be an infinite 0–1-sequence which begins with 1 and is easily possible with respect to P (formally, it is required that $p(\omega) < \infty$). For each positive integer n define k_n as the positive integer whose binary expansion coincides with the first n terms ω^n of ω. By example 1 we have $p(\omega \mid k_n) \succcurlyeq 2^n \asymp k_n p(\omega)$ (remember that ω is fixed and $p(\omega) < \infty$). We see that $p(\omega \mid k)$ may be far from $p(\omega)$.

3.2. *Part II*

The aim of this paper is to present principal ideas rather than to obtain results in their most advanced form, so our setting is as simple as possible. Generalizations in various directions are straightforward (though technically complicated).

We consider a family $(P_\theta\colon \theta \in \Omega)$ of probability distributions in Ω (the set of all infinite 0–1-sequences) with the parameter set again Ω. Fix a probability distribution Q in the parameter set Ω. The set Ω will be interpreted as the binary expansions of real numbers in [0, 1] (working directly with the set [0, 1] would involve some technical difficulties). The distributions T, Y and the family Z were defined in Section 1.

The family (P_θ) is supposed *computable*. This means that the function $P_\theta\{\Gamma(x)\}$ of arguments θ and x is computable. We also suppose that the distribution Q is computable. It is easy to verify that the distributions T and Y are computable as well. Often we shall additionally suppose that (Z_ω) is computable. Standard probability distributions and families are computable.

A lower semicomputable non-negative function $f_\theta(\omega)$ is called a *measure of impossibility with respect to (P_θ)* if

$$\int f_\theta(\omega)\, P_\theta(\mathrm{d}\omega) < 1$$

for all θ. There is a universal (i.e. largest to within a constant factor) measure of impossibility. We fix such a measure p. The value $p_\theta(\omega)$ is called the *level of impossibility of* ω *with respect to* P_θ. Since the constant $\frac{1}{2}$ is a measure of impossibility, $p_\theta(\omega) \geqslant 1$.

Example 2. Let $\omega_0 \in \Omega$ be a fixed computable point, Ω' be the set Ω without ω_0 and $P_\theta(\Omega')$ be a computable function of θ. Then the holder of (P_θ) is morally certain that, if the true value θ of the parameter is such that P_θ is almost entirely concentrated at ω_0, then the outcome ω is ω_0 (equivalently, he believes that $\omega \neq \omega_0$ is almost impossible under such θ). The quantitative form of this assertion is $p_\theta(\omega) \geqslant 1/P_\theta(\Omega')$, where $\omega \neq \omega_0$. The proof is simple: the function of the arguments $\omega \in \Omega$ and θ equal to $1/\{2\,P_\theta(\Omega')\}$, for $\omega \neq \omega_0$, and equal to 0, for $\omega = \omega_0$, is a measure of impossibility.

Example 3. For $\theta \in [0, 1]$ and $\sigma \in [0, 1]$ let $N_{\theta,\sigma}$ be the 'trimmed' normal distribution in [0, 1] with parameters (θ, σ): if ξ is a normal random variable with mean θ and variance σ^2, then $N_{\theta,\sigma}\{0\}$ is the probability that $\xi \leqslant 0$, $N_{\theta,\sigma}\{1\}$ is the probability that $\xi \geqslant 1$ and $N_{\theta,\sigma}(A)$, where A is a measurable set in $]0, 1[$, is the probability that $\xi \in A$. We shall consider θ and σ as ranging over Ω and $N_{\theta,\sigma}$ as a probability distribution in Ω. The *level of impossibility* $n_{\theta,\sigma}(\omega)$ is defined as the largest (to within a constant) factor lower semicomputable non-negative function satisfying $\int n_{\theta,\sigma}\,\mathrm{d}N_{\theta,\sigma} < 1$ for all θ and σ. Let $\sigma \neq 0$ be small. Then outcomes ω many σ from θ are almost impossible, namely

$$n_{\theta,\sigma}(\omega) \geqslant \exp(\tau^2/3),$$

where $\tau = |\omega - \theta|/\sigma$. Indeed, for some $\epsilon > 0$ the function $\epsilon \exp(\tau^2/3)$ is a measure of impossibility.

We also consider conditional measures of impossibility $f_\theta(\omega\,|\,\iota)$, where ι ranges over finite objects (the requirement $\int f_\theta\,\mathrm{d}P_\theta < 1$ should be modified in a natural way: $\int f_\theta(\omega\,|\,\iota)\,P_\theta(\mathrm{d}\omega) < 1$ for all θ and ι). The corresponding level of impossibility is called *conditional* and is denoted by $p_\theta(\omega\,|\,\iota)$. The proof of lemma 6 shows that, if k ranges over positive integers.

$$p_\theta(\omega) \geqslant p_\theta(\omega\,|\,k) \geqslant k^2 p_\theta(\omega). \tag{1}$$

Analogously we can speak of the level of impossibility $t(\theta, \omega)$ with respect to the joint distribution T and $y(\omega)$ with respect to the Bayesian mixture Y. When the posterior family (Z_ω) is computable we introduce the corresponding level of impossibility $z_\omega(\theta)$. By remark 3 we can, and we shall, suppose that the level of impossibility always takes values of the form 2^k, k being an integer.

4. LEVEL OF IMPOSSIBILITY WITH RESPECT TO T AND Y

Recall that $\{P_\theta(\mathrm{d}\omega)\}$ is a computable family and $Q(\mathrm{d}\theta)$ is a computable probability distribution describing beliefs of the individual. The joint distribution $T(\mathrm{d}\theta, \mathrm{d}\omega)$, the Bayesian mixture $Y(\mathrm{d}\theta)$ and the posterior distributions $Z_\theta(\mathrm{d}\theta)$ are defined in Section 1. The level of impossibility with respect to P, Q, T, Y and Z is denoted by p, q, t, y and z. The following theorem is an analogue of theorem 1 of Gacs (1974) due to Levin.

Theorem 1.

$$t(\theta, \omega) \asymp q(\theta)\,p_\theta\{\omega\,|\,q(\theta)\}.$$

Proof for the inequality $\geqslant$. First we shall prove the inequality

$$2^k p_\theta(\omega | 2^k) \preccurlyeq 2^{k'} p_\theta(\omega | 2^{k'}), \tag{2}$$

where k and k' range over the integers such that $k < k'$. Putting $n = k' - k$ and using the inequality

$$p_\theta(\omega | \iota) \preccurlyeq p_\theta(\omega | \iota, n) \preccurlyeq n^2 p_\theta(\omega | \iota),$$

which is analogous to inequality (1), we obtain

$$\begin{aligned} 2^{k'} p_\theta(\omega | 2^{k'}) &\asymp 2^{k'} p_\theta(\omega | k') = 2^{k+n} p_\theta(\omega | k+n) \succcurlyeq 2^k (2^n/n^2) p_\theta(\omega | k+n, n) \\ &\asymp 2^k (2^n/n^2) p_\theta(\omega | k, n) \succcurlyeq 2^k p_\theta(\omega | k) \asymp 2^k p_\theta(\omega | 2^k). \end{aligned}$$

The last inequality follows from $2^n/n^2 \geqslant 1$ and

$$p_\theta(\omega | k, n) \succcurlyeq p_\theta(\omega | k).$$

Denote

$$\phi(\theta, \omega) = \sup_{2^k \leqslant q(\theta)} \{2^k p_\theta(\omega | 2^k)\}.$$

Inequality (2) (with $2^{k'} = q(\theta)$) implies

$$\phi(\theta, \omega) \asymp q(\theta) p_\theta\{\omega | q(\theta)\}.$$

It is evident that the function $\phi(\theta, \omega)$ is lower semicomputable. Applying Fubini's theorem, we obtain

$$\begin{aligned} \int \phi(\theta, \omega)\, T(\mathrm{d}\theta, \mathrm{d}\omega) &\asymp \int q(\theta) p_\theta\{\omega | q(\theta)\}\, T(\mathrm{d}\theta, \mathrm{d}\omega) \\ &= \int \left[\int p_\theta\{\omega | q(\theta)\}\, P_\theta(\mathrm{d}\omega) \right] q(\theta)\, Q(\mathrm{d}\theta) < 1. \end{aligned}$$

So, for some constant $\epsilon > 0$, the function $\epsilon\,\phi(\theta, \omega)$ is a measure of impossibility with respect to T and, therefore,

$$t(\theta, \omega) \succcurlyeq \phi(\theta, \omega) \asymp q(\theta) p_\theta\{\omega | q(\theta)\}.$$

Proof for the inequality $\preccurlyeq$. Let $\phi(\theta) = \int t(\theta, \omega) P_\theta(\mathrm{d}\omega)$. By Fubini's theorem we have

$$\int \phi(\theta)\, Q(\mathrm{d}\theta) = \int t(\theta, \omega)\, T(\mathrm{d}\theta, \mathrm{d}\omega) < 1.$$

Moreover, the function $\phi(\theta)$ is lower semicomputable. Hence

$$\phi(\theta) \preccurlyeq q(\theta). \tag{3}$$

By lemma 3 there is a computable non-decreasing sequence of non-negative simple functions $t_n(\theta, \omega)$ which satisfy $t = \sup(t_n)$. By inequality (3) there is a rational constant c such that $\phi(\theta) < c\,q(\theta)$ for all θ. We put

$$\psi(\theta, \omega, k) = \sup_n \left\{ (ck)^{-1} t_n(\theta, \omega) : (ck)^{-1} \int t_n(\theta, \omega) P_\theta(\mathrm{d}\omega) < 1 \right\}$$

(k is a positive integer). The function $\psi(\theta, \omega, k)$ is lower semicomputable and, for all θ and k,

$$\int \psi(\theta, \omega, k) P_\theta(\mathrm{d}\omega) \leqslant 1.$$

Hence, $p_\theta(\omega \mid k) \succcurlyeq \psi(\theta, \omega, k)$. By the choice of c,

$$\psi(\theta, \omega, k) = (ck)^{-1} t(\theta, \omega)$$

when $k = q(\theta)$. Therefore,

$$p_\theta\{\omega \mid q(\theta)\} \succcurlyeq q(\theta)^{-1} t(\theta, \omega). \qquad \square$$

This theorem and inequality (1) show that, roughly, a pair (θ, ω) agrees with T if and only if θ agrees with Q and ω agrees with P_θ. The equality in theorem 1 can be violated when the condition '$\mid q(\theta)$' is dropped; however, this equality can be replaced by the following double inequality.

Corollary 1. For any fixed $\epsilon > 0$,

$$q(\theta) p_\theta(\omega) \preccurlyeq t(\theta, \omega) \preccurlyeq q^{1+\epsilon}(\theta) p_\theta(\omega).$$

Proof. The left-hand inequality follows from theorem 1 and the inequality $p_\theta(\omega) \preccurlyeq p_\theta\{\omega \mid q(\theta)\}$; the right-hand inequality follows from

$$p_\theta\{\omega \mid q(\theta)\} \asymp p_\theta\{\omega \mid \log_* q(\theta)\} \preccurlyeq p_\theta(\omega)\{\log_* q(\theta)\}^2$$

(see inequality (1); $\log_*$ was defined in remark 4). $\square$

The proof shows that corollary 1, as well as those of the results below that involve ϵ, can be strengthened (further strengthening is provided by remark 4).

Corollary 2. Suppose that the family (Z_ω) is computable. For any fixed $\epsilon > 0$,

$$\frac{\{q(\theta) p_\theta(\omega)\}^{1-\epsilon}}{z_\omega(\theta)} \preccurlyeq y(\omega) \preccurlyeq \frac{\{q(\theta) p_\theta(\omega)\}^{1+\epsilon}}{z_\omega(\theta)}. \tag{4}$$

Proof. Theorem 1 implies

$$y(\omega) \asymp \frac{q(\theta) p_\theta\{\omega \mid q(\theta)\}}{z_\omega\{\theta \mid y(\omega)\}}.$$

It remains to apply inequality (1). $\square$

This corollary demonstrates that $y(\omega)$ and $z_\omega(\theta)$ are small whenever $q(\theta)$ and $p_\theta(\omega)$ are small (the Bayesian method is empirically valid). One more consequence of theorem 1 is the following 'empirical' analogue of Bayes's theorem.

Corollary 3. If the family (Z_ω) is computable, then, for any fixed $\epsilon > 0$,

$$\frac{\{q(\theta) p_\theta(\omega)\}^{1-\epsilon}}{\inf_\theta\{q(\theta) p_\theta(\omega)\}} \preccurlyeq z_\omega(\theta) \preccurlyeq \frac{\{q(\theta) p_\theta(\omega)\}^{1+\epsilon}}{\inf_\theta\{q(\theta) p_\theta(\omega)\}}.$$

Proof. We shall prove only the 'exact' equality

$$z_\omega\{\theta \mid y(\omega)\} \asymp \frac{q(\theta) p_\theta\{\omega \mid q(\theta)\}}{\inf_\theta[q(\theta) p_\theta\{\omega \mid q(\theta)\}]}.$$

By theorem 1 this equality is equivalent to

$$z_\omega\{\theta|y(\omega)\} \asymp t(\theta,\omega)/\inf_\theta\{t(\theta,\omega)\}. \tag{5}$$

Again applying theorem 1 we obtain

$$\inf_\theta\{t(\theta,\omega)\} \asymp \inf_\theta[y(\omega)\,z_\omega\{\theta|y(\omega)\}] \asymp y(\omega), \tag{6}$$

so equation (5) reduces to theorem 1. □

Example 4. What follows is a variant of Lindley's paradox (Shafer, 1982). The family $(N_{\theta,\sigma})$ is defined in example 3. Let us fix a small $\sigma \neq 0$. We consider an individual who ascribes the prior probability $\frac{1}{2}$ to the value $\theta = \theta_0 = 100...$ (which corresponds to $\frac{1}{2}$) and distributes the remaining probability $\frac{1}{2}$ over Ω uniformly. Suppose that he observes an outcome ω such that $\tau = |\omega-\theta_0|/\sigma$ is large but $\exp(\tau^2/2)$ is much smaller than $1/\sigma$. Simple computation yields

$$Z_\omega(\Omega') \sim \sqrt{(2\pi)}\,\sigma\exp(\tau^2/2) \ll 1$$

(Ω' is Ω without θ_0), so the individual is sure that $\theta=\theta_0$ (compare example 2). This seems strange since the outcome ω is many σ from θ_0 (compare example 3).

To be precise, we simultaneously have $z_\omega(\theta) \gg 1$ for $\theta \neq \theta_0$ (i.e. it is morally impossible that $\theta \neq \theta_0$) and $p_{\theta_0}(\omega) \gg 1$ (i.e. ω is morally impossible under the hypothesis $\theta=\theta_0$). These two inequalities contradict each other under the additional assumption that the individual is empirically valid: indeed, corollary 2 shows that $y(\omega)$ is large whenever there is a θ_0 such that $z_\omega(\theta_0)$ is small (i.e. θ_0 is easily possible *a posteriori*) and $p_{\theta_0}(\omega)$ is large. The paradox disappears when we admit that the individual's opinion *may* be wrong; corollary 2 then implies that it *is* wrong (see the discussion in Shafer (1982)). (Note that we have not used the fact that θ_0 is the only parameter value that is easily possible under Z_ω.) Another interesting consequence of corollary 2 is that the individual's opinion will be falsified whenever some *a priori* almost impossible θ (i.e. such that $q(\theta) \gg 1$) becomes easily possible after learning the outcome ω of the experiment (i.e. $z_\omega(\theta)$ is small).

Remark 6. A straightforward probabilistic analysis shows that, in the situation of example 4, the Y probability of obtaining a paradoxical result is small. We thus arrive at the same conclusion: the paradoxical outcome makes us consider that our whole set-up is suspect. An advantage of the algorithmic approach is that this conclusion becomes a manifestation of a general law, namely that expressed by relation (4). Another advantage is the pointwise character of our analysis: relation (4) implies not only that the Y-probability of obtaining a paradoxical outcome is small but also that an outcome ω can be paradoxical only when $y(\omega)$ is large (see remark 2).

Remark 7. In our discussion of Lindley's paradox we have used the fact that the constant c implicitly involved in the inequalities of relation (4) does not depend on σ.

It turns out that the important equality (between the extreme terms of) (6) holds even without the assumption of the computability of (Z_ω).

Theorem 2.

$$y(\omega) \asymp \inf_\theta\{t(\theta,\omega)\}.$$

Proof for the inequality $\leqslant$. Since

$$\int y(\omega)\, T(\mathrm{d}\theta, \mathrm{d}\omega) = \int y(\omega)\, Y(\mathrm{d}\omega) < 1,$$

we have $y(\omega) \leqslant c\, t(\theta, \omega)$ for all θ and ω, where c is some constant. Hence, for all ω,

$$y(\omega) \leqslant c \inf_{\theta}\{t(\theta, \omega)\}.$$

Proof for the inequality $\geqslant$. Note that

$$\int \inf_{\theta}\{t(\theta, \omega)\}\, Y(\mathrm{d}\omega) = \int \inf_{\theta}\{t(\theta, \omega)\}\, T(\mathrm{d}\theta, \mathrm{d}\omega)$$
$$\leqslant \int t(\theta, \omega)\, T(\mathrm{d}\theta, \mathrm{d}\omega) < 1.$$

It remains to prove that the function

$$\phi(\omega) = \inf_{\theta}\{t(\theta, \omega)\}$$

is lower semicomputable (and thus measurable: the measurability of ϕ has been implicitly used above), i.e. that the relation $\phi(\omega) > r$ (r is a rational number) is positively decidable. Let $t_n(\theta, \omega)$ be a non-decreasing sequence of simple functions converging to $t(\theta, \omega)$ (see lemma 3). Designate by $A_n(\omega, r)$ the set of all θ satisfying $t_n(\theta, \omega) > r$; it is a *simple set*, i.e. a finite union of sets of the form $\Gamma(x)$. Then A is a computable function and, for arbitrary fixed ω and r, $A_n(\omega, r)$ is a non-decreasing sequence of simple sets. The inequality $\phi(\omega) > r$ is equivalent to $\cup A_n(\omega, r) = \Omega$. Furthermore, by the compactness of Ω it is equivalent to $\exists\, n$: $A_n(\omega, r) = \Omega$. The last relation is obviously positively decidable. □

Corollary 4. For any fixed $\epsilon > 0$,

$$\inf_{\theta}\{q(\theta) p_{\theta}(\omega)\} \lessdot y(\omega) \lessdot \inf_{\theta}\{q^{1+\epsilon}(\theta) p_{\theta}(\omega)\}.$$

Proof. See corollary 1 and theorem 2. □

5. DISCUSSION

In conclusion we briefly describe our position on the arbitrariness in the definition of level of impossibility. It is true that, though every two universal measures of impossibility coincide to within a constant factor, no one particular outcome can be ascribed a definite level of impossibility: choosing different universal measures of impossibility we can obtain arbitrarily small and arbitrarily large values for the level of impossibility. In this sense the algorithmic approach to statistics is essentially asymptotical. Its use can be justified by the following considerations:

(a) the algorithmic approach can provide a deeper understanding of some controversial issues (see, for example, example 4 earlier) and thereby lead to new developments outside this approach;
(b) future investigations may lead to understanding what universal measures of impossibility are 'reasonable'; if any two reasonable universal measures of impossibility coincide to within a *small* factor, we will be able to choose one

of the measures as the standard (see the end of section 3 of Kolmogorov (1965));

(c) in a particular situation we can use some 'approximations' to the level of impossibility which both satisfy the necessary asymptotic properties of the level of impossibility and are tailored to the situation.

REFERENCES

Dawid, A. P. (1985a) The impossibility of inductive inference. *J. Am. Statist. Ass.*, **80**, 340–341.

——— (1985b) Calibration-based empirical probability (with discussion). *Ann. Statist.*, **13**, 1251–1273.

——— (1986) A Bayesian view of statistical modelling. In *Bayesian Inference and Decision Making* (eds P. Goel and A. Zellner), pp. 391–404. Amsterdam: North-Holland.

de Finetti, B. (1937) La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincare*, **7**, 1–68. (Engl. transl. (1964): Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (eds H. E. Kyburg and H. E. Smokler), pp. 93–158. New York: Wiley.)

Gacs, P. (1974) On the symmetry of algorithmic information. *Sov. Math. Dokl.*, **15**, 1477.

——— (1980) Exact expressions for some randomness tests. *Z. Math. Log. Grndlag. Math.*, **26**, 385–394.

Kolmogorov, A. N. (1965) Three approaches to the quantitative definition of information. *Prob. Informn Transmssn*, **1**, 1–7.

Kolmogorov, A. N. and Uspenskii, V. A. (1987) Algorithms and randomness. *Theory Probab. Applic.*, **32**, 389–412.

Levin, L. A. (1984) Randomness conservation inequalities: information and independence in mathematical theories. *Informn Control*, **61**, 15–37.

Martin-Löf, P. (1966) The definition of random sequences. *Informn Control*, **9**, 602–619.

Popper, K. R. (1934) *Logik der Forschung*. Vienna: Springer. (Engl. transl. (1959): *The Logic of Scientific Discovery*. London: Basic Books.)

Rissanen, J. (1987) Stochastic complexity (with discussion). *J. R. Statist. Soc.* B, **49**, 223–239, 252–265.

Rogers, H. (1967) *Theory of Recursive Functions and Effective Computability*. New York: McGraw-Hill.

Schnorr, C. P. (1977) A survey of the theory of random sequences. In *Basic Problems in Methodology and Linguistics* (eds R. E. Butts and J. Hintikka), pp. 193–211. Dordrecht: Reidel.

Shafer, G. (1982) Lindley's paradox (with discussion). *J. Am. Statist. Ass.*, **77**, 325–351.