JOURNAL OF COMPUTER AND SYSTEM SCIENCES 55, 96–104 (1997)
ARTICLE NO. SS971502

V. G. Vovk†

Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom

Received July 6, 1995; revised September 26, 1996

We consider the problem of learning as much information as possible about the parameter θ of the Bernoulli model $\{P_\theta \mid \theta \in [0,1]\}$ from the statistical data $x \in \{0,1\}^n$, $n \geqslant 1$ being the sample size. Explicating this problem in terms of the Kolmogorov complexity and Rissanen's minimum description length principle, we construct a computable point estimator which (a) extracts from x all information it contains about θ, and (b) discards all sample noise in x. Our result is closely connected with Rissanen's theorem about the optimality of his scheme of coding statistical data.

1. MAIN RESULTS

The *Bernoulli model* is the set $\{P_\theta \mid \theta \in [0,1]\}$ of probability distributions in $\{0,1\}$ defined by $P_\theta\{1\} = \theta \ \forall \theta$. Suppose we are given data $x \in \{0,1\}^n$ (with the sample size $n \in \mathbb{N} := \{1,2,\ldots\}$ known *a priori*), and we believe that for some θ the probability distribution P_θ^n in $\{0,1\}^n$ provides a good description for x. The question that interests us is:

What can we learn from x about θ?

In our simple case of the Bernoulli model, it is possible to answer this question exhaustively: we shall extract from x all and only *useful information* (i.e., information about θ) it contains. In other words, we shall split all the information in x into the useful information and the useless noise. To make this precise, we shall need several definitions.

Let f and g be real-valued functions of arguments $x_1, \ldots, x_n$. Then $f \leqslant^+ g$ (resp. $f \leqslant^\cdot g$) means that there is a constant $c > 0$ such that $f(x_1, \ldots, x_n) \leqslant g(x_1, \ldots, x_n) + c$ (resp. $f(x_1, \ldots, x_n) \leqslant cg(x_1, \ldots, x_n)$) for all $x_1, \ldots, x_n$; the notation $f \leqslant^\cdot g$ will be used only for nonnegative f and g. Instead of $f \leqslant^+ g$ and $f \leqslant^\cdot g$ we shall sometimes write $g \geqslant^+ f$ and $g \geqslant^\cdot f$, respectively; $f =^+ g$ (resp. $f =^\cdot g$) means that both $f \leqslant^+ g$ and $g \leqslant^+ f$ (resp. $f \leqslant^\cdot g$ and $g \leqslant^\cdot f$). Church's λ-notation will often be used: e.g., $\lambda xy.(x+y)$ is the function which transforms x, y into $x+y$; when a prefix like λxy. is clear from the context, we drop it. Expressions like $\lambda x_1 \cdots x_n.A =^+ \lambda x_1 \cdots x_n.B$ are abbreviated to $\lambda x_1 \cdots x_n.A =^+ B$.

First we discuss how we can measure the amount of information. We shall briefly describe Kolmogorov's algorithmic approach as modified by Levin and Chaitin (for details see, e.g., Li and Vitányi [3] or V'yugin [10]). We let $\mathbb{Z}$ stand for the set of integers. A partial function F from $\{0,1\}^* \times \mathbb{Z}$ to $\mathbb{Z}$ is called a *prefix coding scheme* if, for all $p, q \in \{0,1\}^*$ and $n \in \mathbb{Z}$,

$$\left.\begin{array}{r} F(p,n) \text{ is defined} \\ p \text{ is a proper prefix of } q \end{array}\right\} \Rightarrow F(q,n) \text{ is undefined.}$$

(This is a "conditional" variant of the standard notion of information theory.) If, in addition, F is a partial computable function, we call it a *computable prefix coding scheme*. A string $p \in \{0,1\}^*$ is an *F-description* of an integer m given n if $m = F(p,n)$. Let $K_F(m \mid n)$ denote the length of a shortest F-description of m given n. (We shall say that K_F is the *minimum description length function* for F.) There exists a computable prefix coding scheme U such that $\lambda mn.K_U(m \mid n) \leqslant^+ K_F(m \mid n)$, for any other computable prefix coding scheme F (see [3] or [10]). We fix one of such U and call it the *Levin–Chaitin scheme*; $K(m \mid n) := K_U(m \mid n)$ is called the *prefix complexity* of m given n. If θ is a computable real number, we define $K(\theta \mid n)$ to be $\min_m K(m \mid n)$, m ranging over the Gödel numbers of θ (and put $K(\theta \mid n) := \infty$ if θ is not computable). By $K(x \mid n)$, $x \in \{0,1\}^*$, we mean $K(m \mid n)$, where m is the number of x in a fixed enumeration of $\{0,1\}^*$.

Now we know how we can measure the total amount of information in x: since the sample size n is given in advance, a natural measure is $K(x \mid n)$. The next question is: what is the *useful* information in x? The abstract theory of Kolmogorov complexity (Li and Vitányi [3, Chaps. 2 and 3]) does not give a satisfactory answer to this question; the optimal description may hopelessly mix up the useful information with noise. We can speak of usefulness of information only when we have some purpose ("useful" for

* The research described in this publication was made possible in part by Grants #MRS000 and #MRS300 from the International Science Foundation and the Russian Government. It was completed while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences (Stanford, CA, USA). I am grateful for financial support provided by the National Science Foundation (#SES-9022192).

† E-mail: vovk@dcs.rhbnc.ac.uk.

0022-0000/97 $25.00

what?). As we already mentioned, this teleological aspect of the data is captured by our statistical model $\{P_\theta\}$. Since our purpose is to learn something about the parameter θ, by "useful information" we shall mean a value θ for the parameter such that P_θ is a "good description" for x. In spirit of Rissanen's approach, our goal will be to break up the information $K(x|n)$ contained in x into the sum of two parts, the useful information and the noise.

The most difficult question is: what is the noise in the data? Let us first consider the case where it is known that x was generated by P^n, P being a known computable probability distribution in $\{0, 1\}$. In this case, x contains nothing but noise (the useful information is the information about the probability distribution that generated the data, and this distribution is already known to us). Compressing x with, say, the Shannon–Fano code (Li and Vitányi [3, Example 1.29]), we obtain a sequence $C(x) \in \{0, 1\}^*$ of length about $-\log P^n\{x\}$ (log always stands for the binary logarithm in this paper). Formally, we can define a computable function $C: \{0, 1\}^* \to \{0, 1\}^*$ such that

$$\lambda x.\text{length}(C(x)) =^+ -\log P^n\{x\}$$

(n is the length of x) and we can efficiently extract x from n and $C(x)$. We fix such C and regard $C(x)$ as the noise in x. This code is very efficient; it is easy to check that

$$\lambda n.\mathbf{E}_{x \in P^n} 2^{\text{length}(C(x)) - K(x|n)} \leqslant^\cdot 1, \tag{1}$$

where $\mathbf{E}_{x \in P^n}$ means averaging over $x \in P^n$ (and, in general, $x \in Q$ means that x is generated from the probability distribution Q). (It can be shown that $\text{length}(C(x)) - K(x|n)$ is the largest, to within an additive constant, enumerable function that satisfies this property—cf. Li and Vitányi [3, Theorems 4.3 and 4.2].) By Chebyshev's inequality, (1) implies

$$\lambda nm.P^n\{x \in \{0, 1\}^n \mid K(x|n) \leqslant \text{length}(C(x)) - m\} \leqslant^\cdot 2^{-m},$$

so we can be practically sure that the length of our code $C(x)$ for x is close to the length $K(x|n)$ corresponding to the optimal coding scheme.

Now let us return to our family $\{P_\theta\}$. For each parameter value θ we can efficiently encode x using about $-\log P_\theta^n\{x\}$ bits; let $C_\theta(x)$ be the corresponding code-word. Therefore, $C: [0, 1] \times \{0, 1\}^* \to \{0, 1\}^*$ is a computable function such that

$$\lambda\theta x.\text{length}(C_\theta(x)) =^+ -\log P_\theta^n\{x\},$$

and there is a computable function which maps each triple $(n, \theta, C_\theta(x))$, where $x \in \{0, 1\}^n$, into x. (Strictly speaking, $C_\theta(x)$ will depend not only on θ itself but also on the constructive representation of θ; for simplicity, the reader can always assume that θ in $C_\theta(x)$ is computable and is represented by its Gödel number.) As before, we fix C and interpret $C_\theta(x)$ as the noise in x (w.r.t. P_θ).

Now suppose that we are lucky enough to find a value θ for the parameter such that $K(x|n)$ (recall that x is our data of size n) is close to

$$\text{DL}_\theta(x) := -\log P_\theta^n\{x\} + K(\theta|n). \tag{2}$$

This means that we have split the $K(x|n)$ bits of information in x into the $K(\theta|n)$ bits of useful information and the $-\log P_\theta^n\{x\}$ bits of noise. In other words, we have extracted all useful information and only useful information from x.

We define the *statistical coding scheme* as follows. Each code-word for $x \in \{0, 1\}^n$, $n \geqslant 1$, consists of two parts: the preamble, which is a description of some $\theta \in \Theta$ under the Levin–Chaitin scheme, and the body, $C_\theta(x)$. The preamble encodes the useful information, and the body is the noise (we need not compress the noise: being incompressible is part of our understanding of noise). The minimal possible code length under this coding scheme is within an additive constant of

$$\text{DL}(x) := \inf_\theta \text{DL}_\theta(x).$$

The function DL is analogous to Rissanen's [7, 8] stochastic complexity; we shall call it *statistical complexity*. The next theorem and its corollary show that when x is generated by some P_θ, $\text{DL}(x)$ is close to $K(x|n)$ with high probability (in the terminology of Dawid [1], our inference model agrees with the production model).

THEOREM 1 (Separation theorem).

$$\lambda n\theta.\mathbf{E}_{x \in P_\theta^n} 2^{\text{DL}(x) - K(x|n)} \leqslant^\cdot 1. \tag{3}$$

This theorem strengthens inequality (1). It asserts that when x is typical w.r.t. our model $\{P_\theta\}$, the information in x can be split into the useful information (which, by Theorem 2 below, can be efficiently extracted from x) and the noise.

By Jensen's inequality, Theorem 1 implies

COROLLARY 1. $\lambda n\theta.\mathbf{E}_{x \in P_\theta^n} K(x|n) =^+ \mathbf{E}_{x \in P_\theta^n} \text{DL}(x)$.

Simplest examples show that $\text{DL}(x)$ and $K(x|n)$ are very different for some x; e.g., when x is the sequence $0101\cdots$ of alternating 0s and 1s of length n, we have

$$\lambda n.\text{DL}(x) =^+ n, \quad \lambda n.K(x|n) =^+ 0$$

(this sequence is untypical in the highest degree under the Bernoulli model).

Theorem 1 asserts that, with probability close to 1, the statistical coding scheme is as efficient as the Levin–Chaitin scheme; in other words, the information in x can be split into the useful information and the noise. Now we shall consider the question of whether we can do it efficiently.

A *point estimator* is a function E of the type $\{0,1\}^*\backslash\{\Box\} \to [0,1]$ ($\Box \in \{0,1\}^*$ is the empty sequence). It is *computable* if all values $E(x)$, $x \in \{0,1\}^*\backslash\{\Box\}$, are computable, and there exists an algorithm which transforms each $x \in \{0,1\}^*\backslash\{\Box\}$ into a Gödel number of $E(x)$. The next theorem asserts that the useful information in x can be extracted efficiently.

THEOREM 2. *There exists a computable point estimator E such that*

$$\lambda x.\mathrm{DL}(x) =^+ \mathrm{DL}_{E(x)}(x),$$

x ranging over $\{0,1\}^*\backslash\{\Box\}$ (where we recall that $E(x)$ is a parameter value; see (2)).

2. CONNECTIONS WITH RISSANEN'S THEOREM

In this section we discuss connections of our "statistical complexity" $\mathrm{DL}(x)$ with Rissanen's stochastic complexity and connections of our Theorems 1 and 2 with Rissanen's well-known result (Theorem 1 of [6] and [5]; see also [7]). Rissanen's result is very general (it covers even statistical models with a variable number of parameters) but it requires that the model be "regular" in some sense. The full Bernoulli model $\{P_\theta \mid \theta \in [0,1]\}$ is not quite "regular" (say, Fisher's information is infinite at $\theta = 0$ and $\theta = 1$), and we shall only consider its submodel $\{P_\theta \mid \theta \in [\varepsilon, 1-\varepsilon]\}$, where $0 < \varepsilon < \frac{1}{2}$ is a fixed rational constant.

Recall that in Section 1 we defined the statistical coding scheme as encoding each $x \in \{0,1\}^*\backslash\{\Box\}$ by a two-part code; the preamble, which is a description of some θ under the Levin–Chaitin scheme, and the body, $C_\theta(x)$. A straightforward coding scheme provides, given n, each element of the net

$$\Theta_n := [\varepsilon, 1-\varepsilon] \cap \{an^{-1/2} \mid a \in \mathbb{Z}\} \tag{4}$$

with a description of length at most $\lceil \frac{1}{2}\log n \rceil$ (other θ may have no description). Replacing the Levin–Chaitin scheme with this coding scheme in the definition of the statistical coding scheme, we obtain Rissanen's (see, e.g., [6]) coding scheme. The minimum description length function for Rissanen's coding scheme is within an additive constant from

$$\mathrm{DL}^*(x) := \min_{\theta \in \Theta_n} (-\log P^n_\theta\{x\} + \tfrac{1}{2}\log n),$$

n being the length of x. In essence, $\mathrm{DL}^*(x)$ is the stochastic complexity of x. Notice that this function will change by at most an additive constant (and we ignore such a change in this paper) if we replace the uniform discretization scale (4) by the discretization scale proportional to the inverse of the Fisher information (as suggested by Rissanen in [9, p. 57]).

It is easy to see that

$$\lambda x.K(x \mid n) \leqslant^+ \mathrm{DL}(x) \leqslant^+ \mathrm{DL}^*(x), \tag{5}$$

where we recall that

$$\mathrm{DL}(x) = \inf_\theta (-\log P^n_\theta\{x\} + K(\theta \mid n)).$$

In Section 1 (Corollary 1) we saw that the left-hand inequality becomes an equality "on the average":

$$\lambda n\theta.\mathbf{E}_{x \in P^n_\theta} K(x \mid n) =^+ \mathbf{E}_{x \in P^n_\theta} \mathrm{DL}(x).$$

The right-hand inequality of (5) does not hold even "on the average," as our next theorem shows.

THEOREM 3.

$$\lambda n\theta.\mathbf{E}_{x \in P^n_\theta} \mathrm{DL}(x) =^+ H_n(\theta) + K(\theta_n \mid n), \tag{6}$$

$$\lambda n\theta.\mathbf{E}_{x \in P^n_\theta} \mathrm{DL}^*(x) =^+ H_n(\theta) + \tfrac{1}{2}\log n, \tag{7}$$

where θ_n is an element of Θ_n closest to θ and $H_n(\theta)$ is the binary entropy of P^n_θ:

$$H_n(\theta) := \mathbf{E}_{x \in P^n_\theta}(-\log P^n_\theta\{x\}).$$

We have

$$\lambda n\theta.K(\theta_n \mid n) \leqslant^+ \tfrac{1}{2}\log n;$$

besides, when θ has small deficiency of randomness (in the sense of Martin-Löf; see, e.g., V'yugin [10, Section 1]) w.r.t. the uniform probability distribution in $[0,1]$ given n, $K(\theta_n \mid n)$ is close to $\frac{1}{2}\log n$. However, it is easy to find θ for which $K(\theta_n \mid n)$ is very different from $\frac{1}{2}\log n$; say, for $\theta = \frac{1}{2}$ we have $\lambda n.K(\theta_n \mid n) =^+ 0$.

Theorems 1, 2, and 3 immediately imply the following analog of Rissanen's [6, Theorem 1] specialized to $\{P_\theta \mid \theta \in [\varepsilon, 1-\varepsilon]\}$. (An important difference of Rissanen's framework is that n is not assumed to be given in advance in it.)

COROLLARY 2. (a) $\lambda n\theta.\mathbf{E}_{x \in P^n_\theta} K(x \mid n) + \delta(\theta_n \mid \Theta_n) \geqslant^+ H_n(\theta) + \frac{1}{2}\log n$, *where $\delta(\theta_n \mid \Theta_n) := \log |\Theta_n| - K(\theta_n \mid n)$ (notice that $|\Theta_n| =^\cdot n^{1/2}$) is the prefix randomness deficiency of θ_n in Θ_n (cf. Li and Vitányi [3]);*

(b) $\lambda n\theta.\mathbf{E}_{x \in P^n_\theta} \mathrm{DL}^*(x) \leqslant^+ H_n(\theta) + \frac{1}{2}\log n$.

In Section 1 we were preoccupied with how to efficiently find a θ at which the statistical complexity (i.e., the inf in $\inf_\theta \mathrm{DL}_\theta(x)$) is attained to within an additive constant. In the case of stochastic complexity this problem becomes trivial: if E is a computable point estimator such that

$$\forall x\colon E(x) \in \Theta_n,$$

$$k/n \leqslant \varepsilon \Rightarrow E(x) = \inf \Theta_n,$$

$$k/n \geqslant 1-\varepsilon \Rightarrow E(x) = \sup \Theta_n,$$

and, when x ranges so that $\varepsilon \leqslant k/n \leqslant 1-\varepsilon$,

$$\lambda x.\, |E(x) - k/n| \leqslant^{\cdot} n^{-1/2}$$

(n is the length of x and k is the number of 1s in x), then

$$\lambda x.\, \mathrm{DL}^*(x) =^+ -\log P^n_{E(x)}\{x\} + \tfrac{1}{2}\log n.$$

3. CONNECTIONS WITH THE THEORY OF POINT ESTIMATION

In this section we shall informally discuss the properties of the point estimator E from Theorem 2. Usually one considers a point estimator E satisfactory if $E(x)$ is close to θ with high probability or "on the average" when x is generated by the distribution P_θ. In this paper we adopt another measure of performance suggested by Rissanen's minimum description length principle ([4–7]; see also Sections 4 and 6 of Dawid [2] and Section 5.7 of Li and Vitányi [3]). Theorem 2 asserts (in the case of the Bernoulli model) the existence of a point estimator E which satisfies the following desiderata:

- E is computable;
- E extracts all the useful information from x and discards all the noise in x, in the sense that $\mathrm{DL}(x) =^+ \mathrm{DL}_{E(x)}(x)$.

We shall briefly discuss the nature of difficulties we shall have to overcome.

The maximum likelihood estimator $\hat\theta(x) = k/n$, where n is the length of x and k is the number of ones in x, is the optimal, or at least a very good, point estimator under many standard performance criteria. Why is it unsuitable for our purpose? The problem with the maximum likelihood estimator is that the estimates $\hat\theta(x)$ contain too much information (and so part of it is noise). It can be shown that in typical cases we have

$$\mathrm{DL}_{\hat\theta(x)}(x) - \mathrm{DL}(x) \approx \frac{1}{2}\log\frac{k(n-k)}{n}.$$

Another natural approach might be to try to find an argument θ at which $\mathrm{DL}_\theta(x)$ attains its minimum for the given data x. Here the problem is that the function $\lambda\theta.\,\mathrm{DL}_\theta(x)$ is noncomputable. Let us, nevertheless, consider the function

$$\check\theta(x) := \arg\min_\theta \mathrm{DL}_\theta(x).$$

It is instructive to compare the reasons why $\hat\theta$ and $\check\theta$ fall short of satisfying our desiderata. The estimator $\hat\theta$ is easily computable by a very simple algorithm but the estimates $\hat\theta(x)$ are typically too complex (in the sense of Kolmogorov complexity). On the other hand, the estimates $\check\theta(x)$ are simple enough but we can see no way to compute them (although the proof of Theorem 2 will show that for some special choice of the complexity function K even $\check\theta$ is computable).

It is easy to see that the maximum likelihood estimator can be improved (in the sense of our performance criterion). Our purpose is to minimize the sum in (2), and the maximum likelihood estimate minimizes only the first addend. Since near an extremum the changes are small ("Fermat's principle"), we can decrease the sum in (2) reporting not the whole of $\hat\theta(x)$ but only the most significant digits of $\hat\theta(x)$. (So we must balance likelihood and simplicity.) In this paper we in essence use Rissanen's [4–9] optimal truncation scheme which suggests reporting $\hat\theta(x)$ with accuracy $n^{-1/2}$ (this corresponds to reporting the first $\frac{1}{2}\log n$ fractional binary digits of $\hat\theta(x)$). Things are complicated by the fact that the Bernoulli model is not quite "regular": the Fisher information $I(\theta) = 1/\theta(1-\theta)$ tends to infinity as $\theta \to 0$ or $\theta \to 1$. Because of this, we shall have to use a nonuniform net (see Subsection 4.1).

Roughly, only the first $\frac{1}{2}\log(n^3/(k(n-k)))$ bits of the $\log n$ bits of the maximum likelihood estimate k/n are useful information, and the other bits are noise. Removing the

$$\log n - \frac{1}{2}\log\frac{n^3}{k(n-k)} = \frac{1}{2}\log\frac{k(n-k)}{n}$$

bits containing noise, we obtain an estimate satisfying our requirements.

Of course, all our results are far away from the most interesting aspects of Rissanen's theory (such as dealing with parameters of variable dimensions). It seems implausible that we shall be able to extract all and only information about the parameter from the data in the case of more extensive statistical models; therefore, a reasonable goal might be to find out how much information about the parameter we can extract from the data while discarding all (or almost all) noise.

In the conference version of this paper [11] the emphasis was put on constructing computable estimators E such that

$\mathrm{DL}(x) =^+ \mathrm{DL}_{E(x)}(x)$ (they are called "minimum description length estimators" there). In it we also prove the existence of such estimators for two Gaussian models: the family $\{N_{\mu,1} \mid \mu \in \mathbb{R}\}$ of all Gaussian distributions in the real line with variance 1, and the family $\{N_{0,\sigma} \mid \sigma > 0\}$ of all Gaussian distributions in the real line with mean 0 and positive variance.

4. PROOFS

4.1. *Preliminaries*

In this section, x ranges over $\{0,1\}^*\backslash\{\Box\}$; n and k denote the length and the number of ones in x, respectively.

First, we shall describe the estimator E whose existence is asserted in Theorem 2; this estimator will also play a crucial role in our proof of Theorem 1. For each sample size $n = 1, 2, \ldots$, we define a net

$$\{\sin^2(an^{-1/2}) \mid a = 1, \ldots, \lfloor \pi n^{1/2}/2 \rfloor - 1\} \subseteq [0,1]. \quad (8)$$

A rough definition of our estimate $E(x)$ is: $E(x)$ is the element of the net $\sin^2(an^{-1/2})$ closest to k/n. The problem with this definition is that it may be computationally infeasible to find such $E(x)$: when k/n is exactly halfway between two adjacent elements of the net, the algorithm trying to compute $E(x)$ will find itself in the situation of "Buridan's ass." Therefore, we accept the following less natural definition. The point estimator $E: \{0,1\}^*\backslash\{\Box\} \to [0,1]$ is a computable function which, for any $x \in \{0,1\}^*\backslash\{\Box\}$, satisfies the following: If k/n is outside the convex closure of the net, $E(x)$ is the nearest to k/n element of the net. If k/n belongs to the net, $E(x) = k/n$. Otherwise, let θ' and θ'' be two adjacent elements of the net such that $k/n \in [\theta', \theta'']$. If $|k/n - \theta'|$ and $|k/n - \theta''|$ differ by more than 1%, $E(x)$ is the element of $\{\theta', \theta''\}$ closest to k/n; otherwise, we only require that $E(x) \in \{\theta', \theta''\}$. Without loss of generality we assume that $E(x)$ depends on x only through n and k; we shall sometimes use the notation $E_n(k)$ instead of $E(x)$.

Net (8) might at first look not very natural, but actually it is an implementation of a well-known idea due to Rissanen (see also Yamanishi [12]): a suitable discretization scale is that proportional to the inverse of the Fisher information; (8) is essentially the only net satisfying this property (cf. Corollary 3 below).

LEMMA 1. *When* $n \in \mathbb{N}$, $\alpha \in [1/2, \pi n^{1/2}/2 - 1/2]$, *and* $a, b \in [0, \pi n^{1/2}/2]$ *range so that* $a \leqslant \alpha \leqslant b$ *and* $1/2 \leqslant b - a \leqslant 2$, *we have*

$$\lambda n \alpha b . \sin^2(bn^{-1/2}) - \sin^2(an^{-1/2}) =^\cdot n^{-1/2} \sin(\alpha n^{-1/2}) \cos(\alpha n^{-1/2}). \quad (9)$$

Proof. Equivalent transformations of (9) yield

$$\begin{aligned}
&(\sin(bn^{-1/2}) - \sin(an^{-1/2}))(\sin(bn^{-1/2}) + \sin(an^{-1/2})) \\
&\quad =^\cdot n^{-1/2} \sin(\alpha n^{-1/2}) \cos(\alpha n^{-1/2}); \\
&\cos\left(\frac{b+a}{2} n^{-1/2}\right) \sin\left(\frac{b-a}{2} n^{-1/2}\right) \\
&\qquad \times \sin\left(\frac{b+a}{2} n^{-1/2}\right) \cos\left(\frac{b-a}{2} n^{-1/2}\right) \\
&\quad =^\cdot n^{-1/2} \sin(\alpha n^{-1/2}) \cos(\alpha n^{-1/2}).
\end{aligned}$$

Our task has reduced to proving that

$$\cos\left(\frac{b+a}{2} n^{-1/2}\right) =^\cdot \cos(\alpha n^{-1/2}), \quad (10)$$

$$\sin\left(\frac{b-a}{2} n^{-1/2}\right) =^\cdot n^{-1/2}, \quad (11)$$

$$\sin\left(\frac{b+a}{2} n^{-1/2}\right) =^\cdot \sin(\alpha n^{-1/2}), \quad (12)$$

$$\cos\left(\frac{b-a}{2} n^{-1/2}\right) =^\cdot 1. \quad (13)$$

Equalities (11) and (13) immediately follow from $1/2 \leqslant b - a \leqslant 2$; (10) and (12) reduce, in view of $a \leqslant \alpha \leqslant b$ and $\alpha \in [1/2, \pi n^{1/2}/2 - 1/2]$, to

$$\begin{aligned}
\cos\left(\frac{\pi}{2} - \frac{3}{2} n^{-1/2}\right) &=^\cdot \cos\left(\frac{\pi}{2} - \frac{1}{2} n^{-1/2}\right) \\
&=^\cdot \cos\left(\frac{\pi}{2} - \frac{1}{4} n^{-1/2}\right)
\end{aligned}$$

and

$$\sin\left(\frac{3}{2} n^{-1/2}\right) =^\cdot \sin\left(\frac{1}{2} n^{-1/2}\right) =^\cdot \sin\left(\frac{1}{4} n^{-1/2}\right),$$

respectively; these two relations are equivalent, and the second of them is obviously true. ∎

COROLLARY 3. *When* $k \in \{1, \ldots, n-1\}$,

$$\lambda n k . \# E_n^{-1}(E_n(k)) =^\cdot \sqrt{k(n-k)/n}. \quad (14)$$

Proof. For simplicity, we shall assume that $E_n^{-1}(E_n(k))$ always consists of consecutive elements of the set $\{0, \ldots, n\}$. Define a, α, b by the conditions

$$\sin^2(an^{-1/2}) = \frac{1}{n}\inf E_n^{-1}(E_n(k)),$$

$$\sin^2(bn^{-1/2}) = \frac{1}{n}\sup E_n^{-1}(E_n(k)),$$

$$\sin^2(\alpha n^{-1/2}) = k/n.$$

We can see that $a \leqslant \alpha \leqslant b$, $\alpha \in [1, \pi n^{1/2}/2 - 1]$, and $1/2 \leqslant b - a \leqslant 2$. Since

$$\# E_n^{-1}(E_n(k)) =\cdot n(\sin^2(bn^{-1/2}) - \sin^2(an^{-1/2}))$$

and

$$\begin{aligned}\sqrt{k(n-k)/n} &= \sqrt{n(k/n)(1-k/n)} \\ &= \sqrt{n\sin^2(\alpha n^{-1/2})\cos^2(\alpha n^{-1/2})} \\ &= \sqrt{n}\sin(\alpha n^{-1/2})\cos(\alpha n^{-1/2}),\end{aligned}$$

we can rewrite (14) as

$$\begin{aligned}&\sin^2(bn^{-1/2}) - \sin^2(an^{-1/2}) \\ &\quad =\cdot n^{-1/2}\sin(\alpha n^{-1/2})\cos(\alpha n^{-1/2}),\end{aligned}$$

which coincides with (9). ∎

The following two lemmas describe important properties of the log-likelihood function for the Bernoulli model. We use the notation G_x for the log-likelihood function expressed through the variable a (which will no longer be assumed to be an integer) introduced by $\theta = \sin^2(an^{-1/2})$,

$$G_x(a) := \ln(\sin^{2k}(an^{-1/2})\cos^{2(n-k)}(an^{-1/2})),$$

a ranging over $[0, \pi n^{1/2}/2]$. Since $G_x(a)$ depends on x only through n and k, we shall also use the notation $G_{n,k}(a)$ for $G_x(a)$. We use the notation $\hat{a}(x)$, or $\hat{a}(n, k)$, for the maximum likelihood estimate of the parameter a:

$$\hat{a}(x) := \arg\max_a G_x(a)$$

(therefore, $\sin^2(\hat{a}(n,k)\, n^{-1/2}) = k/n$). We shall often drop n.

LEMMA 2. *When* $n \geqslant 1$, $a \in [1, \pi n^{1/2}/2 - 1]$, *and* $k \in \{1, ..., n-1\}$ *range so that* $|a - \hat{a}(n,k)| < 1$,

$$\lambda ank.\, G_{n,k}(a) =^+ G_{n,k}(\hat{a}(n,k)).$$

Proof. It suffices to prove that the values

$$\sup_a \left|\frac{d^2 G_k(a)}{da^2}\right|, \tag{15}$$

where a ranges over

$$[1, \pi n^{1/2}/2 - 1] \cap [\hat{a}(k) - 1, \hat{a}(k) + 1],$$

do not exceed some bound. Calculating the second derivative, we rewrite (15) as

$$2\sup_a\left(\frac{k/n}{\sin^2(an^{-1/2})} + \frac{1 - k/n}{\cos^2(an^{-1/2})}\right).$$

Note that

$$k/n = \sin^2(\hat{a}(k)\, n^{-1/2}), \quad 1 - k/n = \cos^2(\hat{a}(k)\, n^{-1/2}),$$

so it suffices to prove that

$$\sup_a \frac{\sin^2((a+1)\, n^{-1/2})}{\sin^2(an^{-1/2})}, \quad \sup_a \frac{\cos^2((a-1)\, n^{-1/2})}{\cos^2(an^{-1/2})}$$

are bounded above by some constant. It is easy to see that both suprema are equal to

$$\frac{\sin^2(2n^{-1/2})}{\sin^2(n^{-1/2})} \to 4 \quad (n \to \infty),$$

so they are indeed bounded. ∎

LEMMA 3. *Let* $n \geqslant 1$, $a \in [0, \pi n^{1/2}/2]$, *and* $k \in \{1, ..., n-1\}$. *For some constant* $\varepsilon > 0$,

$$\lambda ank.\, G_{n,k}(\hat{a}(n,k)) - G_{n,k}(a) \geqslant^+ \varepsilon\, |a - \hat{a}(n,k)|.$$

Proof. By the symmetry of the problem, we can (and shall) suppose $a > \hat{a}(k)$. Furthermore, we can consider only the case $a \geqslant \hat{a}(k) + \frac{1}{2}$. Since $G_k''(a)$ is negative everywhere, it is sufficient to prove that $-G_k'(\hat{a}(k) + \frac{1}{2})$ is greater than some constant $\varepsilon > 0$. We find

$$-G_k'(a) = 2n^{-1/2}\left((n-k)\frac{\sin(an^{-1/2})}{\cos(an^{-1/2})} - k\frac{\cos(an^{-1/2})}{\sin(an^{-1/2})}\right),$$

so we are required to prove

$$\begin{aligned}&(n-k)\sin^2((\hat{a}+1/2)\, n^{-1/2}) - k\cos^2((\hat{a}+1/2)\, n^{-1/2}) \\ &\quad > \frac{\varepsilon}{2} n^{1/2}\sin((\hat{a}+1/2)\, n^{-1/2})\cos((\hat{a}+1/2)\, n^{-1/2}),\end{aligned}$$

where $\hat{a} := \hat{a}(k)$. This inequality is equivalent to

$$\begin{aligned}&n\sin^2((\hat{a}+1/2)\, n^{-1/2}) - k \\ &\quad > \frac{\varepsilon}{2} n^{1/2}\sin((\hat{a}+1/2)\, n^{-1/2})\cos((\hat{a}+1/2)\, n^{-1/2}),\end{aligned}$$

or

$$\sin^2((\hat{a}+1/2)\,n^{-1/2})-\sin^2(\hat{a}n^{-1/2})$$
$$>\frac{\varepsilon}{2}n^{-1/2}\sin((\hat{a}+1/2)\,n^{-1/2})\cos((\hat{a}+1/2)\,n^{-1/2}).$$

The last inequality immediately follows from Lemma 1. ∎

4.2. *Proof of Theorem* 1

We can rewrite (3) as the conjunction of three inequalities,

$$\lambda n\theta.\mathbf{E}_{x\in P_\theta^n}2^{\mathrm{DL}(x)-K(x|n)}\mathbf{I}(0<k<n)\leqslant^{\cdot}1, \tag{16}$$
$$\lambda n\theta.2^{\mathrm{DL}(0\cdots0)-K(0\cdots0|n)}P_\theta^n\{0\cdots0\}\leqslant^{\cdot}1, \tag{17}$$
$$\lambda n\theta.2^{\mathrm{DL}(1\cdots1)-K(1\cdots1|n)}P_\theta^n\{1\cdots1\}\leqslant^{\cdot}1 \tag{18}$$

($\mathbf{I}(A)$ is 1 if property A holds, and 0, if not).

Inequalities (17) and (18) immediately follow from

$$\mathrm{DL}(0\cdots0)=^+K(0\cdots0|n)=^+\mathrm{DL}(1\cdots1)$$
$$=^+K(1\cdots1|n)=^+0,$$

so we are only required to prove (16). We shall prove

$$\mathbf{E}_{x\in P_\theta^n}2^{-\log P^n_{E(x)}\{x\}+K(E(x)|n)-K(x|n)}\mathbf{I}(0<k<n)\leqslant^{\cdot}1. \tag{19}$$

Let $\{0,1\}^n_k$ stand for the set of all $x\in\{0,1\}^n$ that contain exactly k ones. Since

$$\mathbf{E}_{x\in P_\theta^n}\mathbf{I}(0<k<n)=\mathbf{E}_{k\in\mathrm{bin}_{n,\theta}}\mathbf{I}(0<k<n)\,\mathbf{E}_{x\in\{0,1\}^n_k}$$

(where $\mathrm{bin}_{n,\theta}$ is the binomial distribution in $\{0,...,n\}$ with parameter θ, $\mathrm{bin}_{n,\theta}\{k\}:=\binom{n}{k}\theta^k(1-\theta)^{n-k}$, and the subscript $x\in\{0,1\}^n_k$ implies that x are chosen from the uniform distribution in $\{0,1\}^n_k$) and

$$\mathbf{E}_{x\in\{0,1\}^n_k}2^{-K(x|n)}=\frac{1}{\binom{n}{k}}\sum_{x\in\{0,1\}^n_k}2^{-K(x|n)}=^{\cdot}\frac{1}{\binom{n}{k}}2^{-K(k|n)}$$

(the right-hand equality follows from the coincidence of 2^{-K} with the *a priori* probability; see V'yugin [10] or Li and Vitányi [3]), we can rewrite (19) as

$$\mathbf{E}_{k\in\mathrm{bin}_{n,\theta}}\mathbf{I}(0<k<n)\frac{1}{\binom{n}{k}}2^{-\log P^n_{E(x)}\{x\}+K(E(x)|n)-K(k|n)}\leqslant^{\cdot}1$$

(remember that $P^n_{E(x)}\{x\}$ and $E(x)$ depend on x only through n and k); i.e.,

$$\mathbf{E}_{k\in\mathrm{bin}_{n,\theta}}\mathbf{I}(0<k<n)$$
$$\times2^{(-\log\binom{n}{k}-\log P^n_{E(x)}\{x\})+(K(E(x)|n)-K(k|n))}\leqslant^{\cdot}1. \tag{20}$$

By Stirling's formula we find

$$-\log\binom{n}{k}-\log P^n_{E(x)}\{x\}$$
$$=^+-\frac{1}{2}\log n-n\log\frac{n}{e}+\frac{1}{2}\log k+k\log\frac{k}{e}$$
$$+\frac{1}{2}\log(n-k)+(n-k)\log\frac{n-k}{e}$$
$$-k\log E(x)-(n-k)\log(1-E(x))$$
$$=\frac{1}{2}\log\frac{k(n-k)}{n}+k\log\frac{k}{n}+(n-k)\log\frac{n-k}{n}$$
$$-k\log E_n(k)-(n-k)\log(1-E_n(k)).$$

By Lemma 2, we further obtain

$$-\log\binom{n}{k}-\log P^n_{E(x)}\{x\}=^+\frac{1}{2}\log\frac{k(n-k)}{n};$$

therefore, (20) reduces to

$$\mathbf{E}_{k\in\mathrm{bin}_{n,\theta}}\mathbf{I}(0<k<n)\,2^{(1/2)\log(k(n-k)/n)+K(E_n(k)|n)-K(k|n)}\leqslant^{\cdot}1. \tag{21}$$

Since

$$\mathbf{E}_{k\in\mathrm{bin}_{n,\theta}}=\mathbf{E}_{\omega\in E_n\{0,...,n\}}\mathbf{E}_{k\in E_n^{-1}(\omega)}$$

(where $\omega\in E_n\{0,...,n\}$ is distributed in accordance with the image distribution $E_n(\mathrm{bin}_{n,\theta})$ and $k\in E_n^{-1}(\omega)$ is distributed in accordance with the conditional distribution $\mathrm{bin}_{n,\theta}|E_n(k)=\omega$), (21) reduces to

$$\mathbf{E}_{k\in E_n^{-1}(\omega)}\mathbf{I}(0<k<n)\sqrt{(k(n-k)/n)}\,2^{K(\omega|n)-K(k|n)}\leqslant^{\cdot}1$$

(ω ranging over $E_n\{0,...,n\}$). The last inequality reduces, by Corollary 3, to

$$\sum_{k\in E_n^{-1}(\omega)}2^{K(\omega|n)-K(k|n)}\leqslant^{\cdot}1,$$

or, equivalently,

$$\sum_{k\in E_n^{-1}(\omega)}2^{-K(k|n)}\leqslant^{\cdot}2^{-K(\omega|n)}.$$

This inequality follows from the coincidence of 2^{-K} with the *a priori* probability. ▮

4.3. *Proof of Theorem* 2

We are required to prove $\mathrm{DL}(x) =^+ \mathrm{DL}_{E(x)}(x)$. In the case $k=0$, this reduces to

$$0 =^+ -\log((1-\sin^2 n^{-1/2})^n),$$

which is easy to validate. By the symmetry, $\mathrm{DL}(x) =^+ \mathrm{DL}_{E(x)}(x)$ is also true when $k=n$. Therefore, we assume $k \in \{1, ..., n-1\}$.

Our goal is to prove

$$\lambda\theta nk.\ -\log \Lambda_{n,k}(E_n(k)) + K(E_n(k)\,|\,n) \leqslant^+ -\log \Lambda_{n,k}(\theta) + K(\theta\,|\,n), \tag{22}$$

where $\Lambda_{n,k}(\theta) := \theta^k(1-\theta)^{n-k}$ is the likelihood function and k ranges over $\{1, ..., n-1\}$. It is easy to see that

$$\begin{aligned} K(E_n(k)\,|\,n) &=^+ K(\lfloor \hat{a}(n,k) \rfloor\,|\,n) \\ &\leqslant^+ K(\lfloor \hat{a}(n,k) \rfloor, \theta\,|\,n) \\ &\leqslant^+ K(\theta\,|\,n) + K(\lfloor \hat{a}(n,k) \rfloor\,|\,n, \theta), \end{aligned}$$

so to ensure (22) it suffices to prove

$$-\log \Lambda_{n,k}(E_n(k)) + K(\lfloor \hat{a}(n,k) \rfloor\,|\,n, \theta) \leqslant^+ -\log \Lambda_{n,k}(\theta).$$

Expressing θ through a, $\theta = \sin^2(an^{-1/2})$, and using Lemma 2, we reduce this to

$$G_{n,k}(a) + K(\lfloor \hat{a}(n,k) \rfloor\,|\,n, a) \ln 2 \leqslant^+ G_{n,k}(\hat{a}(n,k)).$$

Lemma 3 shows that it suffices to prove

$$K(\lfloor \hat{a}(n,k) \rfloor\,|\,n, a) \ln 2 \leqslant^+ \varepsilon\,|a - \hat{a}(n,k)|.$$

We find

$$\begin{aligned} K(\lfloor \hat{a}(n,k) \rfloor\,|\,n, a) \ln 2 &\leqslant^+ K(\lfloor \hat{a}(n,k) \rfloor\,|\,\lfloor a \rfloor) \ln 2 \\ &\leqslant^+ K(\lfloor \hat{a}(n,k) \rfloor - \lfloor a \rfloor) \ln 2 \\ &\leqslant^+ \varepsilon |\lfloor \hat{a}(n,k) \rfloor - \lfloor a \rfloor| \\ &=^+ \varepsilon\,|\hat{a}(n,k) - a|. \end{aligned}$$

4.4. *Proof of Theorem* 3

Recall that we now assume $\theta \in [\varepsilon, 1-\varepsilon]$.

LEMMA 4.

$$\lambda n\theta.\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} \left(-k \log \frac{k}{n} - (n-k) \log \frac{n-k}{n} \right) =^+ H_n(\theta). \tag{23}$$

Proof. We can rewrite (23) as

$$\begin{aligned} n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} &\left(-\frac{k}{n} \log \frac{k}{n} - \frac{n-k}{n} \log \frac{n-k}{n} \right) \\ &=^+ n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} \left(-\frac{k}{n} \log \theta - \frac{n-k}{n} \log(1-\theta) \right), \end{aligned}$$

or, letting $K_\theta(\xi)$ stand for the Kullbach distance,

$$\xi \log \frac{\xi}{\theta} + (1-\xi) \log \frac{1-\xi}{1-\theta},$$

between the probability distributions $(\xi, 1-\xi)$ and $(\theta, 1-\theta)$,

$$n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} K_\theta(k/n) =^+ 0.$$

We shall separately prove that

$$n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} K_\theta(k/n)\, \mathbf{I}(k/n \in [\varepsilon/2, 1-\varepsilon/2]) =^+ 0 \tag{24}$$

and

$$n \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} K_\theta(k/n)\, \mathbf{I}(k/n \notin [\varepsilon/2, 1-\varepsilon/2]) =^+ 0. \tag{25}$$

Taylor's formula reduces (24) to

$$n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} \tfrac{1}{2} K''_\theta(\xi)(k/n-\theta)^2\, \mathbf{I}(k/n \in [\varepsilon/2, 1-\varepsilon/2]) =^+ 0,$$

where ξ is between θ and k/n. Noting that

$$\begin{aligned} n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} &\tfrac{1}{2} K''_\theta(\xi)(k/n-\theta)^2\, \mathbf{I}(k/n \in [\varepsilon/2, 1-\varepsilon/2]) \\ &\leqslant n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} \tfrac{1}{2} K''_\theta(\xi)(k/n-\theta)^2\, \mathbf{I}(\xi \in [\varepsilon/2, 1-\varepsilon/2]) \\ &\leqslant\cdot\, n\, \mathbf{E}_{k \in \mathrm{bin}_{n,\theta}} (k/n-\theta)^2 = \theta(1-\theta) =^+ 0 \end{aligned}$$

completes the proof of (24).

To prove (25), note that the functions K_θ, $\varepsilon \leqslant \theta \leqslant 1-\varepsilon$, are uniformly bounded, and so (25) reduces to

$$n\, \mathrm{bin}_{n,\theta}\{k\,|\,k/n \notin [\varepsilon/2, 1-\varepsilon/2]\} \leqslant\cdot 1.$$

The last inequality follows from

$$\mathrm{bin}_{n,\varepsilon}\{k\,|\,k/n \in [0, \varepsilon/2]\} \leqslant\cdot 1/n,$$

which is a special case of the standard large-deviation inequalities. ▮

To prove equality (6) of Theorem 3 we rewrite it, using Theorem 2, as

$$\mathbf{E}_{x\in P_\theta^n}\mathrm{DL}_{E(x)}(x) =^+ H_n(\theta) + K(\theta_n|n),$$

i.e.,

$$\mathbf{E}_{x\in P_\theta^n}(-k\log E(x) - (n-k)\log(1-E(x)) + K(E(x)|n)) \\ =^+ H_n(\theta) + K(\theta_n|n). \tag{26}$$

By Lemma 2, we can replace the first two occurrences of $E(x)$ by k/n (Lemma 2 does not cover the case $k\in\{0,n\}$, which is simple); therefore, Lemma 4 reduces (26) to

$$\mathbf{E}_{x\in P_\theta^n}K(E(x)|n) =^+ K(\theta_n|n). \tag{27}$$

It is easy to see that

$$|K(E(x)|n) - K(\theta_n|n)| \leqslant^+ |E(x)-\theta_n|\, n^{1/2}$$

(since, given n and one of the values θ_n or $E(x)$, we can restore the other if we know the number of elements of the net between θ_n and $E(x)$), so (27) follows from

$$\mathbf{E}_{x\in P_\theta^n}|E(x)-\theta_n|\, n^{1/2} =^+ \mathbf{E}_{x\in P_\theta^n}|k/n-\theta|\, n^{1/2} \\ = \mathbf{E}_{x\in P_\theta^n}\left|\frac{k-\theta n}{\sqrt{n\theta(1-\theta)}}\right|\sqrt{\theta(1-\theta)} \\ =^+ 0$$

(the last equality is a direct consequence of the uniform central limit theorem).

Equality (7) of Theorem 3 immediately follows from Lemma 2 and Lemma 4.

ACKNOWLEDGMENTS

This paper profited greatly from Jorma Rissanen's questions and comments on an early draft. The two referees' comments are gratefully appreciated.

REFERENCES

1. A. P. Dawid, Fisherian inference in likelihood and prequential frames of reference (with discussion), *J. R. Statist. Soc. B* **53** (1991), 79–109.
2. A. P. Dawid, Prequential analysis, stochastic complexity and Bayesian inference, *in* "Bayesian Statistics 4" (J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, Eds.), Oxford Univ. Press, Oxford, 1992.
3. M. Li and P. Vitányi, "An Introduction to Kolmogorov Complexity and Its Applications," Springer-Verlag, New York, 1993.
4. J. Rissanen, A universal prior for integers and estimation by minimum description length, *Ann. Statist.* **11** (1983), 416–431.
5. J. Rissanen, Universal coding, information, prediction, and estimation, *IEEE Trans. Inform. Theory* **30** (1984), 629–636.
6. J. Rissanen, Minimum-description-length principle, *in* "Encyclopedia of Statistical Sciences" (S. Kotz and N. L. Johnson, Eds.), Vol. 5, pp. 523–527, Wiley, New York, 1985.
7. J. Rissanen, Stochastic complexity and modeling, *Ann. Statist.* **14** (1986), 1080–1100.
8. J. Rissanen, Stochastic complexity (with discussion), *J. R. Statist. Soc. B* **49** (1987), 223–239; 252–265.
9. J. Rissanen, "Stochastic Complexity in Statistical Inquiry," World Scientific, Singapore, 1989.
10. V. V. V'yugin, Algorithmic entropy (complexity) of finite objects and its applications to defining randomness and amount of information, *Selecta Math. Soviet.* **13** (1994), 357–389.
11. V. G. Vovk, Minimum description length estimators under the optimal coding scheme, *in* "Computational Learning Theory" (P. Vitányi, Ed.), pp. 237–251, Lecture Notes in Computer Science, Vol. 904, Springer-Verlag, Berlin, 1995.
12. K. Yamanishi, A learning criterion for stochastic rules, *Mach. Learning* **9** (1992), 165–203.