Universal Forecasting Algorithms

V. G. Vovk

Research Council for Cybernetics, Academy of Sciences,
40 ulitsa Vavilova, Moscow 117333, USSR

We consider forecasting systems which, when given an initial segment of a binary string, guess the next bit (deterministic forecasting) or estimate the probability of the next bit being 1 (probabilistic forecasting). The quality of forecasting is measured by the number of errors (in the deterministic case) or by the sum of distances between the forecasts and the "true" probabilities (in the probabilistic case). A forecasting system is said to be simple if it has a short description (e.g., a simple formula) which admits fast computation of the forecasts. There is a "universal forecasting algorithm" which for any given bound \( f \) on time of forecasting computes forecasts within time \( f \), and the quality of these forecasts is not much lower than that of any simple forecasting system (the complexity of the "rival" forecasting system may increase as \( f \) increases). The aim of the paper is to study the possibilities and limitations of universal forecasting.

1. Notation and Conventions

\( \mathbb{B} \) is the set \( \{0, 1\} \), \( \mathbb{B}^n \) is \( \bigcup_{i \leq n} \mathbb{B}^i \), and \( \mathbb{B}^\infty \) is \( \bigcup_{i \leq \infty} \mathbb{B}^i \). Elements of \( \mathbb{B}^* \) are called strings. The length of \( p \in \mathbb{B}^* \) is denoted by \( l(p) \). For \( x, y \in \mathbb{B}^* \), \( y \preceq x \) means that \( y \) is an initial segment (or prefix) of \( x \). The sets of real, rational, and integer numbers are designated by \( \mathbb{R}, \mathbb{Q}, \) and \( \mathbb{Z} \), respectively; \( \mathbb{D} \) symbolizes the set \( \{2^{-n} | n = 2, 3, ...\} \). The maximum and minimum of two numbers are denoted by \( \vee \) and \( \wedge \), respectively. We set \( t^+ = 0 \vee t \), \( t^\oplus = 1 \wedge t \), where \( t \in \mathbb{R} \), and \( T^+ = T \cap [0, \infty[ \), \( T^\oplus = T \cap [1, \infty[ \), where \( T \subset \mathbb{R} \). Natural and binary logarithms are denoted by \( \ln \) and \( \log \), respectively. \( \lfloor t \rfloor \) and \( \lceil t \rceil \) stand for

\[
\max\{n \in \mathbb{Z} | n \leq t\}, \min\{n \in \mathbb{Z} | n \geq t\},
\]

respectively, \( t \) being a real number. \( \mathbb{P} \) and \( \mathbb{E} \) are the symbols of probability and expectation, respectively.

For short we write "polynomial" meaning a polynomial of one variable with positive integer coefficients. Constants are positive integer until explicitly stated otherwise. Writing that \( A(v) \) holds for all \( v \) we mean that \( A(v) \) holds for all \( v \) such that \( A(v) \) is defined.

"Algorithm" may be regarded as "Turing machine." Our considerations...
will be invariant under the polynomial transformations of the time axis, so all the results will be machine-independent. Let us fix a sufficiently wide space \( \mathcal{F} \) of finite objects containing the empty object. We restrict our attention to algorithms \( \mathcal{A} \) which, receiving as input \( \varphi \in \mathcal{F} \) and getting access to an oracle for \( g = g_0 g_1 \cdots \in \mathbb{B}^\infty \) (this oracle answers questions about the values \( g_i \) for specified \( i \)), output either nothing or \( \mathcal{A}^g(\varphi) \in \mathcal{F} \).

Let \( \nu: B \to \mathcal{O} \) for some \( B \subset \mathbb{B}^\infty \); \( \nu \) is called a \textit{system of notation for (objects in) } \( \mathcal{O} \) if \( \nu(B) = \mathcal{O} \). We define an algorithm \( \mathcal{A} \) (and the corresponding computable function) to be \textit{\( \nu \)-invariant} if

\[
\nu(g) = \nu(h) \Rightarrow \mathcal{A}^g(\varphi) \simeq \mathcal{A}^h(\varphi),
\]

where \( a \simeq b \) means that either both \( a \) and \( b \) are undefined or both are defined and equal. We define \( \nu \) to be \textit{negative} if such an algorithm \( \mathcal{B} \) exists that, for all \( g \in \nu^{-1}(\mathcal{O}) \) and \( h \in \mathbb{B}^\infty \),

\[
\mathcal{B}^{g,h} \text{ converges} \iff h \notin \nu^{-1}(\mathcal{O}) \text{ or } \nu(g) \neq \nu(h)
\]

(the superscript \( g, h \) may be understood as the sequence \( g_0 h_0 g_1 h_1 g_2 \cdots \)).

**Lemma 1.** Let \( B \subset \mathbb{B}^\infty \) and \( \nu: B \to \mathcal{O} \) be negative. There exists a computable function, transforming any Gödel number of any computable function of the type \( \mathcal{F} \times \mathbb{B}^\infty \to \mathcal{F} \) to a Gödel number of a \( \nu \)-invariant computable function of the same type, such that any Gödel number of any \( \nu \)-invariant function is transformed to a Gödel number of the same function.

**Proof.** Almost immediately follows from König's lemma.

We fix a negative system of notation \( \nu \) for a sufficiently large \( \mathcal{O} \). For \( \nu \)-invariant algorithms \( \mathcal{A} \) we use the self-evident notation \( \mathcal{A}^i(\varphi) \), where \( i \in \mathcal{O} \) and \( \varphi \in \mathcal{F} \). An oracle for \( i \in \mathcal{O} \) is an oracle for some \( g \in \nu^{-1}(i) \).

### 2. Probabilistic Forecasting I

We formulate the problem of probabilistic forecasting as follows. Some device sequentially generates \( n \) bits composing a string \( x = x_0 x_1 \cdots x_{n-1} \). Our aim is to estimate the probability of the event \( x_i = 1 \) given an initial segment \( x^i = x_0 x_1 \cdots x_{i-1} \) \((i < n)\) of \( x \). A variant of this problem is central in the “prequential approach” to the foundations of statistics (Dawid, 1984).

Let \( P(x) \) be the probability of generating a string \( x \in \mathbb{B}^n \). Then the “behavior” of the device is determined by the function \( P: \mathbb{B}^n \to \mathbb{R}^+ \), which satisfies the condition

\[
\sum_{x \in \mathbb{B}^n} P(x) = 1.
\]
Such functions will be called \textit{n-measures}. We use the notation

\begin{align*}
P(A) &= \sum_{x \in A} P(x), \quad P(y) = \sum_{y \in x} P(x), \\
P(j\mid y) &= \begin{cases} 
P(yj)/P(y), & \text{if } P(y) \neq 0, \\
1/2, & \text{otherwise,}
\end{cases}
\end{align*}

where $A \subseteq \mathbb{B}^n$, $y \in \mathbb{B}^<\!n$, and $j \in \mathbb{B}$.

When solving our problem some restrictions are usually imposed on the \textit{n-measure} $P$. In statistics it is often supposed that $P$ is known a priori except for several numerical parameters. In recent time weaker "non-parametric" assumptions about $P$ are gaining popularity, but using such assumptions deteriorates quality of forecasting as compared to the "parametric" case. Besides, in statistics one usually supposes the independence and identical distribution of the "coordinate" random variables. R. J. Solomonoff (1978) supposes the computability of $P$ (he considers devices generating infinite sequences). In the author's work (Vovk, 1989) the simplicity of $P$ is supposed. Now we want to construct "universal" forecasting algorithms which do not depend on assumptions concerning $P$ like these. It should be noted that Solomonoff's (1964, 1978) forecasting methods are universal (in our sense), though noncomputable.

Our aim is to construct a forecasting system which estimates $P(1\mid x')$, the probability of $x_i = 1$ given $x'$. An arbitrary function $\mu$, which transforms each $y \in \mathbb{B}^<\!n$ to a rational number $\mu(y) \in [0, 1]$ (an estimate of the probability of the next bit being 1), is called an \textit{n-forecasting system}. As an example of an \textit{n-forecasting system} may serve Laplace's rule of succession

\begin{equation}
\hat{\lambda}(y) = \frac{k + 1}{|l(y) + 2|}, \quad y \in \mathbb{B}^<\!n,
\end{equation}

where $k$ is the number of 1's in $y$.

It seems reasonable to confine one's attention to the class of "simple" forecasting systems. Following (Kolmogorov, 1965, Sect. 4) and (Hartmanis, 1983) we require a forecasting system considered simple to have a short description $p$ (such a description may be, e.g., a simple formula or computer program) such that the computation of the values of the forecasting system given $p$ is fast.

Let $\alpha, \gamma \in \mathbb{R}^+$. An $(\alpha, \gamma)$-\textit{description} of an \textit{n-forecasting system} $\mu$ with respect to an algorithm $\mathfrak{A}$ is by definition an arbitrary string $p \in \mathbb{B}^<\!\alpha$ such that $\mathfrak{A}$, when fed with $n$, $p$, and any $y \in \mathbb{B}^<\!n$, computes (a fraction representing) $\mu(y)$ within time $\gamma$. Forecasting systems having $(\alpha, \gamma)$-descriptions with respect to $\mathfrak{A}$ will be called $(\alpha, \gamma)$-\textit{simple} with respect to $\mathfrak{A}$. As usual, it exists a "best" (with some accuracy) algorithm $\mathfrak{A}$. 
LEMMA 2. There exists an algorithm $\mathcal{U}$ satisfying the following. For any algorithm $\mathcal{A}$ one can find such a constant $c$ and a polynomial $f$ that for all $n, x, y \in \mathbb{Z}^\oplus$ all $n$-forecasting systems that are $(x, y)$-simple with respect to $\mathcal{A}$ are $(x + c, f(y))$-simple with respect to $\mathcal{U}$.

Proof. Let $\{\mathcal{A}_i | i \in \mathbb{Z}^\oplus\}$ be a computable numbering of all algorithms and $1^I_0$ denote the all-1 string of length $i$ extended by adding the bit 0 to the right. On inputs $n, 1^I_0p$, and $y$, $\mathcal{U}$ simulates (with polynomial delay) the computation of $\mathcal{A}_i$ on inputs $n, p$, and $y$.

We fix such an algorithm $\mathcal{U}$; in the sequel "$(x, y)$-simple" is understood as "$(x, y)$-simple with respect to $\mathcal{U}$.”

Let a string $x \in \mathbb{B}^n$, generated in accordance with a measure $P$, be forecast by a forecasting system $\mu$. To evaluate the performance of $\mu$ we use the error of forecasting

$$q_\mu(x, P) = \sum_{i=0}^{n-1} \rho(P(1|x^i), \mu(x^i));$$

the function $\rho: [0, 1]^2 \to \mathbb{R}$ (assessing the difference between the forecast $\mu(x^i)$ and the true value $P(1|x^i)$ at a step $i$) is defined by the equality

$$\rho(u, v) = u \log(u/v) + (1 - u) \log((1 - u)/(1 - v))$$

(we set $\log 0 = 0$, $0/0 = 0$, and $a/0 = \infty$ if $a \neq 0$). This is the distance of Kullback–Leibler between distributions $(u, 1-u)$ and $(v, 1-v)$ (when this distance is defined, $\log$ is usually used rather than $\log$). It is nonnegative and equal to 0 only if $u = v$.

In this section the quality of the forecasting system $\mu$ will be measured by the mean value

$$\Delta_\mu(P) = \sum_{x \in \mathbb{B}^n} P(x) q_\mu(x, P)$$

of $q_\mu(x, P)$ with respect to the “true” measure $P$. This approach (in the spirit of (Solomonoff, 1978)) was suggested by the referee who also contributed the next theorem assessing the quality of a “universal forecasting algorithm” $\mathcal{B}$. For any bound $\Gamma$ on time of forecasting, this algorithm computes forecasts within time $\Gamma$, and the quality of the corresponding forecasting system is not much lower than the quality of any simple (relative to $\Gamma$) forecasting system.

THEOREM 1. There exist an algorithm $\mathcal{B}$, a constant $c$, and a polynomial $f$ such that the following holds. Let $n, \Gamma \in \mathbb{Z}^\oplus$. For $\Gamma \geq c$ all the values

$$\pi(y) = \mathcal{B}(n, \Gamma, y), \quad y \in \mathbb{B}^{<n},$$
are computed within time $\tau$. The function $\pi$ so defined is an $n$-forecasting system. If $\alpha, \gamma \in \mathbb{Z}^\oplus$ satisfy the inequality $f(n2^\alpha\gamma) \leq \tau$, $P$ is an $n$-measure, and $\mu$ is an $(\alpha, \gamma)$-simple $n$-forecasting system, then

$$A_\pi(P) \leq A_\mu(P) + \alpha + 1.01 \ln \alpha + c. \quad (2)$$

Let us see in what sense $\Psi$ is "universal." Let $\mu$ be a forecasting system used, e.g., in statistics. Usually $\mu$ is written down as a simple formula such as (1) ($\alpha$ is a small constant) which permits fast computation of $\mu$ ($\gamma$ is not too large). Therefore, the minimal time $f(n2^\alpha\gamma)$ of computing the forecasts by $\Psi$ is not too large. Inequality (2) shows that the mean error $A_\pi(P)$ of forecasting by $\Psi$ is not much larger than the mean error $A_\mu(P)$ of forecasting by $\mu$.

The next theorem asserts the necessity of the addend $\alpha$ in (2).

**Theorem 2.** There are a constant $c$ and a polynomial $f$ satisfying the following. Let $n \in \mathbb{Z}^\oplus$, $\pi$ be an $n$-forecasting system, and $\alpha \in \mathbb{Z}^\oplus$ do not exceed $n$. One can find an $(\alpha + c, f(n))$-simple $n$-forecasting system $\mu$ and an $n$-measure $P$ such that

$$A_\pi(P) \geq A_\mu(P) + \alpha.$$

Now we investigate running time of the algorithm $\Psi$. The question about restrictions under which polynomial-time forecasting is possible seems to be most important. Theorem 1 shows that $\Psi$ is polynomial-time if one confines oneself to $(\alpha, \gamma)$-simple $n$-forecasting systems with

$$\alpha \leq c \ln n, \quad (3)$$

$$\gamma \leq f(n), \quad (4)$$

where $c$ and $f$ are a fixed constant and a fixed polynomial, respectively. Condition (4) seems to be natural. Is (3) essential? Assume that one-way functions (see below) exist. It turns out that allowing $\alpha$ of order $n^\delta$, where $\delta > 0$ is arbitrarily small, makes polynomial-time forecasting impossible. Now we give necessary definitions following mainly (Goldreich *et al.*, 1986).

Let $l \in \mathbb{Z}^\oplus$ and $f_1$ be a polynomial. A *polynomial-time statistical test* (for strings) is a probabilistic polynomial-time algorithm $T$ that takes as input $f_1(k)$ strings in $\mathbb{B}^k$, $k \in \mathbb{Z}^\oplus$, and outputs either 0 or 1. Let a function $G$ be defined on a set $\bigcup_k D_k$, where $D_k \subset \mathbb{B}^k$, $k \in \mathbb{Z}^\oplus$, and transform each $D_k$ into $\mathbb{B}^k$. They say that $G$ passes the test $T$ if, for any polynomial $f_2$, for all sufficiently large $k$,

$$|p_k^G - p_k| < 1/f_2(k),$$

where $p_k^G$ denotes the probability that $T$ outputs 1 on the strings $G(x_i)$, $i = 1, \ldots, f_1(k)$, where $x_i$ are randomly selected from the uniform distribu-
tion on $\mathbb{B}^k$, and $p_k$ denotes the probability that $\mathcal{T}$ outputs 1 on $f_1(k)$ strings randomly selected from the uniform distribution on $\mathbb{B}^k$. The set $\bigcup_k D_k$ is called sampleable if there is a probabilistic polynomial-time algorithm that, given as input $k$ presented in unary, outputs $x \in D_k$ with probability $1/\text{Card} D_k$. The function $G$ is called a CSB generator if

(a) it is polynomial-time computable;
(b) it passes all polynomial-time statistical tests;
(c) the set $\bigcup_k D_k$ is samplable.

Strings output by CSB generators are called CSB strings.

Let a function $f$ be defined on a set $\bigcup_k D_k$, $D_k \subset \mathbb{B}^k$, and transform each $D_k$ into $D_k$. Let $f^i$ denote $f$ applied $i$ times. Then $f$ is a one-way function if

(a) $f$ is polynomial-time computable;
(b) $f$ is hard to invert; that is, for every probabilistic polynomial-time algorithm $\mathfrak{A}$ and for sufficiently large $k$, for every $1 \leq i \leq k^3$, $\mathfrak{A}(x) \notin f^{-1}(x)$ with probability separated from 0 when $x = f^i(y)$ and $y \in D_k$ is selected with equal probabilities;
(c) $\bigcup_k D_k$ is samplable.

Many functions are currently suspected to be one-way. L. A. Levin has proved the following result that strengthens an earlier result by A. C. Yao (1982).

**THEOREM 3 (Levin, 1985).** There exists a CSB generator if and only if there exists a one-way function.

For every $n$-measure $Q$ the equality

$$\mu(y) = Q(1 \mid y), \quad y \in \mathbb{B}^<n,$$

determines an $n$-forecasting system $\mu$, and vice versa, every $n$-forecasting system $\mu$ determines such an $n$-measure $Q$ that this equality holds provided $Q(y) \neq 0$. We identify $n$-measures with the corresponding $n$-forecasting systems. The $n$-measure corresponding to an $n$-forecasting system $\mu$ is denoted by $\mu'$, or, when no ambiguity arises, simply by $\mu$ (thus $'$ is a projection of the set of all $n$-forecasting systems onto its subset consisting of $n$-measures).

**THEOREM 4.** Suppose one-way functions exist. Let a constant $\delta > 0$ be arbitrarily small. There is a polynomial $f$ satisfying the following. Let an algorithm $\mathfrak{Q}$ be such that the values

$$\pi_n(y) = \mathfrak{Q}(n, y), \quad n \in \mathbb{Z}^\oplus, y \in \mathbb{B}^<n,$$
are computed within polynomial time, and the functions \( \pi_n \) are \( n \)-forecasting systems. Then for infinitely many \( n \) there is an \( (n^\delta, f(n)) \)-simple \( n \)-measure \( P \) such that

\[
A_{\pi_n}(P) \geq n - \delta. \tag{5}
\]

For the \( n \)-forecasting system \( \pi_n \) identically equal to \( \frac{1}{2} \) we have

\[
A_{\pi_n}(P) \leq n \rho(0, \frac{1}{2}) = n.
\]

Comparison with (5) shows that forecasting by \( \mathcal{Q} \) is extremely poor, though an \( (n^\delta, f(n)) \)-simple forecasting system (viz., \( P \) itself) performs perfectly.

3. PROOFS OF THEOREMS 1, 2, 4

In the next lemma the "universal forecasting algorithm" \( \mathcal{B} \) is constructed. The method of its construction is well known (Levin, 1984). To put it crudely, if allowed time of computation is \( \Gamma \), then \( \pi \) is the forecasting system, corresponding to a weighted mean of the measures corresponding to simple forecasting systems.

**Lemma 3.** There exist an algorithm \( \mathcal{B} \), a constant \( c \), and a polynomial \( f \) such that the following holds. Let \( n, \Gamma \in \mathbb{Z}^\oplus \). Then for \( \Gamma \geq c \) the values

\[
\pi(y) = \mathcal{B}(n, \Gamma, y) \in \mathcal{Q}, \quad y \in \mathbb{B}^{\leq n},
\]

are computed within time \( \Gamma \) and the function \( \pi \) is an \( n \)-forecasting system. If \( \alpha, \gamma \in \mathbb{Z}^\oplus \) satisfy the inequality

\[
f(n2^{x_\alpha}) \leq \Gamma \tag{6}
\]

and an \( n \)-forecasting system \( \mu \) is \( (\alpha, \gamma) \)-simple, then

\[
\pi'(x) \geq \mu'(x)2^{-x - 1.01 \log \alpha - c}, \quad \forall x \in \mathbb{B}^n. \tag{7}
\]

**Proof.** A constant \( \delta \in \mathbb{D} \) (small) will be specified in the sequel. \( \mathcal{B}(n, \Gamma, y) \) is defined as follows.

(i) Set \( s = 1 \).

(ii) Let \( p \) be the binary expansion of \( s \) without the leading 1. For \( i = 0, 1, \ldots, l(y) \) set

\[
v^i = \mathcal{U}(n, p, y^i),
\]

if the computation of the last value is performed within time \( \Gamma^\delta/n \) and yields a rational number in \([0, 1]\), and set \( v^i = \frac{1}{2} \), otherwise. Then for \( i = 0, 1, \ldots, l(y) - 1 \) set

\[
u_i = \begin{cases} v^i, & \text{if } y_i = 1, \\ 1 - v^i, & \text{otherwise}, \end{cases}
\]
and set \( u_{i(y)}^s = v_{i(y)}^s \). Compute the two values

\[
A_s = \prod_{i=0}^{l(y)-1} u_i^s, \quad B_s = \prod_{i=0}^{l(y)} u_i^s.
\]

(iii) If \( s < L^\delta \), set \( s = s + 1 \) and go to (ii).

(iv) Output the ratio \( B/A \) as a result, where

\[
B = T^{-1} \sum_{s=1}^{L^\delta} (s[l \log_b s]^{101}) \sum_{s=1}^{L^\delta} (s[l \log_b s]^{101})^{-1} B_s,
\]

\[
A = T^{-1} \sum_{s=1}^{L^\delta} (s[l \log_b s]^{101}) \sum_{s=1}^{L^\delta} (s[l \log_b s]^{101})^{-1} A_s,
\]

\[
T = \sum_{s=1}^{L^\delta} (s[l \log_b s]^{101})^{-1}.
\]

It is easy to verify that step (ii) is performed within time

\[
f_1(n(\Gamma^\delta/n) \log s) = f_1(\Gamma^\delta \log s)
\]

\((f_1, f_2, \ldots \text{ are polynomials})\text{ and the total time of the computation does not exceed}

\[
f_2(\Gamma^{2\delta} \log(\Gamma^\delta)) \leq f_3(\Gamma^\delta).
\]

Now we choose \( \delta \) so that \( f_3(\Gamma^\delta) \leq \Gamma \) will hold for sufficiently large \( \Gamma \).

It remains to verify (7). It follows from (iv) that \( \pi' \) is a weighted mean of the measures corresponding to \( s = 1, \ldots, L^\delta \). Choosing sufficiently large \( f \), we get \( 2^{x+1} \leq \Gamma^\delta \) from (6). Hence among \( s \) there is a number \( S \) whose binary expansion is \( 1p \), where \( p \) is an \((x, y)\)-description of the forecasting system \( \mu \). Once more increasing (if necessary) the polynomial \( f \), ensure that the inequality \( \gamma \leq \Gamma^\delta/n \) holds under (6). Then all the values \( v_i^S \) will coincide with \( \mu(y) \), and the sum determining \( \pi'(x) \) will contain \( \mu'(x) \) with the weight

\[
(TS[l \log_b S]^{101})^{-1} \geq T_0^{-1} 2^{x-\alpha} \log_b S^{-1} 101 \log_b S^{-1}.
\]

where

\[
T_0 = \sum_{s=1}^{\infty} (s[l \log_b s]^{101})^{-1}.
\]

Note that always

\[
A_\mu(P) = \sum_{x \in B^\alpha} \log(P(x)/\mu'(x)) P(x)
\]

(Solomonoff, 1978).
Proof of Theorem 1. Inequality (7) implies that
\[ \log(P(x)/\pi'(x)) \leq \log(P(x)/\mu'(x)) + x + 1.01 \log x + c. \]
Averaging with respect to \( P \) and using (8), we get (2).

For \( z \in \mathbb{B}^n \) let \( V_z \) stand for the \( n \)-measure such that \( V_z(z) = 1 \).

Proof of Theorem 2. For \( y \in \mathbb{B}^x \) let \( \mu_y \) be the \( n \)-forecasting system defined by
\[ \mu_y(x') = \begin{cases} y_i, & \text{if } i < l(y), \\ \frac{1}{2}, & \text{otherwise}, \end{cases} \]
x \in \mathbb{B}^n. Of course, all the \( \mu_y \) are \((\alpha + c, f(n))-simple \) (for some fixed \( c \) and \( f \)). It is sufficient to show the existence of \( y \) and \( x \) such that \( \sigma(x) \geq \mu_y(x) 2^{-x} \). Suppose that such \( y \) and \( x \) do not exist, i.e., \( \pi(x) > \mu_y(x) 2^{-x} \) for all \( x \in \mathbb{B}^n \) and \( y \in \mathbb{B}^x \). Summing over \( x \) such that \( y \subset x \), we get \( \pi(y) > 2^{-x} \). Finally, summing over \( y \) yields \( 1 > 1 \), a contradiction.

Proof of Theorem 4. Choose an integer \( l > 1/\delta \). Let \( n \) be of the form \( k^l \), \( k \in \mathbb{Z}^\oplus \). Suppose \( \delta \in \mathbb{Q} \); this does not restrict generality. We define a polynomial-time statistical test \( \mathcal{T} \) as follows. Receiving as input a string \( z \in \mathbb{B}^n \), \( n = k^l \), \( \mathcal{T} \) outputs 1 if \( q_{n_\delta}(z, V_z) \geq n - \delta \) and 0 otherwise. First we assume that \( z \) is generated in accordance with the uniform \( n \)-measure \( L \). Then
\[ L\{z \mid \mathcal{T}(z) = 0\} = L\{z \mid -\log \pi'_n(z) < n - \delta\} = 2^{-n} \cdot \text{Card}\{z \mid \pi'_n(z) > 2^{-n+\delta}\} < 2^{-\delta}. \]
So, \( \mathcal{T} \) outputs 1 with probability separated from 0. Due to Theorem 3 there exist CSB generators. Now we assume that \( z \) is an output of a CSB generator fed with a string randomly selected from the uniform \( k \)-measure. In this case \( \mathcal{T} \) also outputs 1 with probability separated from 0. Hence a CSB string \( z \in \mathbb{B}^n \) exists such that (5) holds, where \( P = V_z \). It is easy to verify that \( P \) is an \((n^{1/d} + c, f(n))-simple \) \( n \)-measure.

4. Deterministic Forecasting

An \( n \)-forecasting system is called deterministic if it takes values in \( \mathbb{B} \). The quality of forecasting a string \( x \in \mathbb{B}^n \) by a deterministic forecasting system \( \mu \) will be measured by the total number of errors
\[ q_\mu(x) = \text{Card}\{i \mid \mu(x_i) \neq x_i\}. \]
Dealing with a version of the problem of deterministic forecasting where $x$ is an infinite sequence, J. M. Barzdin' and R. V. Freivalds (1972) assume that $x$ is computable (moreover, that $x$ belongs to a known recursively enumerable class of computable sequences). We again avoid such assumptions; it seems more natural to impose requirements of effective computability on the forecaster rather than on the "nature." Analogs of Theorems 1, 2, and 4 for the deterministic case are as follows.

**Theorem 5.** There exist an algorithm $\mathcal{B}$, a constant $c$, and a polynomial $f$ such that the following holds. Let $n, \Gamma \in \mathbb{Z}^\oplus$. For $\Gamma \geq c$ all the values $\pi(y) = \mathcal{B}(n, \Gamma, y), \quad y \in \mathbb{B}^{<n},$

are computed within time $\Gamma$ and $\pi$ is a deterministic $n$-forecasting system. If $\alpha, \gamma \in \mathbb{Z}^\oplus$ satisfy $f(n2^\alpha \gamma) \leq \Gamma$, $x \in \mathbb{B}^n$, and $\mu$ is an $(\alpha, \gamma)$-simple deterministic $n$-forecasting system, then

$$q_\mu(x) \leq 3.52q_\mu(x) + 1.23\alpha + c. \quad (9)$$

**Proof.** Let a constant $\delta \in \mathbb{D}$ be sufficiently small. For $s = 1, 2, \ldots, \lfloor \Gamma^\delta \rfloor$ let $p_s$ be the binary expansion of $s$ without the leading 1, and $\mu_s$ be the $n$-forecasting system such that $\mu_s(y)$, where $y \in \mathbb{B}^{<n}$, equals $\mathcal{U}(n, p_s, y)$, if this value is computed within time $\Gamma^\delta/n$ and lies in $\mathbb{B}$, and 0, otherwise.

$\mathcal{B}(n, \Gamma, y)$ is defined as follows. For $s = 1, 2, \ldots, \lfloor \Gamma^\delta \rfloor$ assign the weight

$$w_s(y) = 2^a 2^{-\lfloor \log_2 \frac{w_{s-1}(y)}{q_\mu(x)} \rfloor} C^q$$

to $\mu_s$, where $a = l(p_s)$, $q = q_\mu(x)$ and $C \in \mathbb{Q} \cap [0, 1]$ will be chosen in the sequel. Set

$$\mathcal{B}(n, \Gamma, y) = \begin{cases} 1, & \text{if } \sum_s w_s(y) \mu_s(y) > 2^{-1} \sum_s w_s(y), \\ 0, & \text{otherwise} \end{cases}$$

(the forecasting systems "vote").

Note that when $\mathcal{B}$ errs (i.e., $\mathcal{B}(n, \Gamma, x') \neq x'$), the total weight of the erring forecasting systems is at least half the total weight of all the forecasting systems. The weight of an erring forecasting system $\mu_s$ is multiplied by $C$ (i.e., $w_s(x'+1) = Cw_s(x')$), so the total weight of all the forecasting systems is multiplied by at most $(1 + C)/2$. Since the terminal value of the total weight is at least $2^{-\frac{\alpha}{2}} 2^{-\frac{\log_2 C}{2}}$, where $q = q_\mu(x)$, the number $N$ of the errors made by $\mathcal{B}$ satisfies

$$c_1((1 + C)/2)^N \geq 2^{-\frac{\alpha}{2}} 2^{-\frac{\log_2 C}{2}}.$$
(c₁ being the initial weight of all the forecasting systems), whence

\[ N \leq (\alpha + 2 \lfloor \log \alpha \rfloor - q \log C + \log c₁) / (1 - \log(1 + C)). \tag{10} \]

Taking \( C = 0.137 \), we get (9).

**Note 1.** The method we made use of in this proof has been independently designed by N. Littlestone and M. K. Warmuth (1989) who call it the Weighted Majority Algorithm. In their work one can find several interesting developments.

**Theorem 6.** There exist such a constant \( c \) and a polynomial \( f \) that the following holds. Let \( n \in \mathbb{Z}^+ \), \( \pi \) be a deterministic \( n \)-forecasting system, and \( \eta, \alpha \in \mathbb{Z}^+ \) not exceed \( n/10 \). One can find an \((\alpha + c, f(n))\)-simple deterministic \( n \)-forecasting system \( \mu \) and a string \( x \in \mathbb{B}^n \) such that

\[
q_\mu(x) \leq \eta + 2, \tag{11}
\]

\[
q_\pi(x) \geq 2\eta + \alpha + 4.99(\eta/3 \land \alpha/12). \tag{12}
\]

**Proof.** Our construction is similar to that used to prove Proposition 9 in (Cohen et al., 1985). We shall use the existence of the Golay code (MacWilliams and Sloane, 1977), a set of \( 2^{12} \) strings (codewords) in \( \mathbb{B}^{23} \) with covering radius 3 (the latter means that each string in \( \mathbb{B}^{23} \) can be transformed into a codeword by changing at most 3 bits). First suppose that \( \eta \) and \( \alpha \) are divisible by 3 and 12, respectively. Denote

\[ m = 2\eta + \alpha + 5(\eta/3 \land \alpha/12), \quad m' = m - \lfloor 0.01(\eta/3 \land \alpha/12) \rfloor. \]

We sequentially define \( x_i, \ i = 0, 1, ..., m' - 1 \), by the equality

\[ x_i = \begin{cases} 1, & \text{if } \pi(x_i) = 0, \\ 0, & \text{otherwise.} \end{cases} \]

Let \( x_i = j \) for \( i \geq m' \), where \( j \) will be defined below. We have \( q_\pi(x) \geq m' \). Fix a one-to-one correspondence between the Golay code and \( \mathbb{B}^{12} \); the string corresponding to a codeword will be referred to as the name of that codeword. For \( k = 0, 1, ..., \lfloor n/23 \rfloor - 1 \), let \( x_{23k}, ..., x_{23k + 22} \) be a codeword which coincides with \( x_{23k}, ..., x_{23k + 22} \) in all but at most three positions, and \( p_k \) be the name of this codeword. Consider 2 cases.

(i) \( \eta \geq \alpha/4 \). Then \( m = 2\eta + (17/12)\alpha \). The \( \mu \) is defined by the equalities

\[
\mu(y^i) = x_i, \quad i = 0, 1, ..., (23/12)\alpha - \lfloor \alpha/1200 \rfloor - 1,
\]

\[
\mu(y^i) = j, \quad i = (23/12)\alpha - \lfloor \alpha/1200 \rfloor, ..., m' - 1,
\]

\[
\mu(y^i) = j, \quad i = m', ..., n - 1,
\]
(as $\frac{23}{12}x \leq m$, the second line is correct), where $y$ ranges over $\mathbb{B}^n$ and $j$ is the bit which is in the majority in the subsequence $(x_{(23/12)x - \lfloor x/1200 \rfloor}, ..., x_{m'-1})$. We have

$$q_\mu(x) \leq 3(x/12) + (m - (23/12)x)/2 = \eta.$$ 

As a description of $\mu$ we may take $p_0 p_1 \cdots p_{(x/12)-1} j$. Hence $\mu$ is $(x+c, f(n))$-simple (under proper choice of $c$ and $f$).

(ii) $\eta < x/4$. Then $m = \frac{11}{3} \eta + x$. Now $\mu$ is defined by

$$\mu(y^i) = x^i, \quad i = 0, 1, ..., (23/3)\eta - \lfloor \eta/300 \rfloor - 1,$$

$$\mu(y^i) = x^i, \quad i = (23/3)\eta - \lfloor \eta/300 \rfloor, ..., m' - 1,$$

$$\mu(y^i) = j, \quad i = m', ..., n - 1,$$

(the correctness of the second line follows from $\frac{23}{3} \eta \leq m$), where $j = 1$ and $y$ ranges over $\mathbb{B}^n$. Again $q_\mu(x) \leq 3(\eta/3) = \eta$. Let $a$ be the binary expansion of $\eta$, and $1^{l(a)}$ be the all-1 string of length $l(a)$. Taking

$$1^{l(a)}0ap_0 \cdots p_{\eta/3 - \lfloor \eta/6900 \rfloor - 1} x_{(23/3)\eta - \lfloor \eta/300 \rfloor} \cdots x_{m'-1}$$

as a description of $\mu$ (the prefix $1^{l(a)}0a$ is needed to enable the separation of the part $p_0 \cdots p_{\eta/3 - \lfloor \eta/6900 \rfloor - 1}$), we see that, for some constants $c_1$ and $c_2$, $\mu$ is $(\alpha' + c_1, f(n))$-simple, where

$$\alpha' = 2l(a) + 1 + 12(\eta/3) - 12\lfloor \eta/6900 \rfloor + (m - (23/3)\eta)$$

$$= \alpha + 2l(a) + 1 - 12\lfloor \eta/6900 \rfloor \leq \alpha + c_2.$$ 

So, when $\alpha$ and $\eta$ are divisible by 12 and 3, respectively, we have $q_\mu(x) \leq \eta$ and (12). In the general case these relations hold with $\alpha$ and $\eta$ replaced by $12\lceil \alpha/12 \rceil$ and $3\lceil \eta/3 \rceil$ respectively, whence (11) and (12) readily follow.

It is easy to see that the only nontrivial term in the right-hand side of inequality (12) is $4.99(\eta/3 \wedge \alpha/12)$. Indeed, for any deterministic $n$-forecasting system $\pi$ there is a string $x$ for which it fails on each of the first $2\eta + \alpha$ bits; let $x_{2\eta + \alpha}, ..., x_{\eta - 1}$ be 1. We can take as $\mu$ a forecasting system which "knows" $x_0, ..., x_{\eta - 1}$ and issues a constant forecast for $x_{\eta}, ..., x_{2\eta + \alpha - 1}$ failing on no more than half of these bits so that $q_\mu(x) \leq \eta$.

Note 2. Let $q_\mu(x)$ approximately equal $\alpha/4$. Then (9) shows that the algorithm $\Psi$ makes at most approximately $2.11\alpha$ errors. On the other hand, (12) shows that every forecasting algorithm sometimes makes at least approximately $1.91\alpha$ errors.
Note 3. In the right-hand side of (10), \( \alpha \) is taken with the weight \( \frac{1}{1 - \log(1 + C)} \), which tends to 1 as \( C \to 0 \), and \( q_{\mu}(x) \) is taken with the weight \( \frac{-\log C}{1 - \log(1 + C)} \), which tends to 2 as \( C \to 1 \). Therefore, in (12) the coefficients 1 of \( t_1 \) and 2 of \( \eta \) cannot be improved. Our choice of \( C \) is nearly optimal when \( q_{\cdot}(x) \) is near to \( \alpha/4 \). When it is expected that the performance of some very simple forecasting system will be not very bad, \( C \) close to 1 should be chosen; on the other hand, if we expect that some not very complex forecasting system performs very well, \( C \) should be close to 0.

We identify a string \( x \in \mathbb{B}^n \) with the deterministic \( n \)-forecasting system \( \mu \) such that \( \mu(y) = x_i, \ i = 0, 1, \ldots, n - 1, \ y \in \mathbb{B}^i \).

**Theorem 7.** Suppose one-way functions exist. Let \( \delta > 0 \) be arbitrarily small and \( c \) arbitrarily large. There is such a polynomial \( f \) that the following holds. Let an algorithm \( \mathcal{Q} \) compute the values

\[
\pi_n(y) = \mathcal{Q}(n, y), \quad n \in \mathbb{Z}^+, \ y \in \mathbb{B} < n,
\]

within polynomial time and the functions \( \pi_n \) be deterministic \( n \)-forecasting systems. Then for infinitely many \( n \) there exists such an \((n^\delta, f(n))-simple\) string \( x \in \mathbb{B}^n \) that

\[
q_{\pi_n}(x) \geq n/2 + c(n \log n)^{1/2}.
\]

**Proof.** Choose an integer \( l > 1/\delta \). Let \( n \) be of the form \( k^l, \ k \in \mathbb{Z}^+ \). We may suppose that \( c \) is an integer. When given \( x \in \mathbb{B}^n \), a polynomial-time statistical test \( \mathcal{I} \) outputs 1 if (13) holds, and 0 otherwise. Assume that \( x \) is generated in accordance with the uniform \( n \)-measure. The large deviation theorem (Feller, 1970, Chap. VII, (6.7)) shows that the probability of (13) is at least \( f^{-1}(n) \) for some polynomial \( f \). Let \( x \) be generated by a CSB generator fed with strings randomly selected from the uniform distribution on \( \mathbb{B}^k \). For sufficiently large \( k \) the probability of (13) does not equal 0. So (13) holds for some CSB string \( x \).

Inequality (13) shows that the forecasts issued by \( \pi \) are more often wrong than right, though an \((n^\delta, f(n))-simple\) forecasting system is never wrong. Note that deleting the term \( c(n \log n)^{1/2} \) in the right-hand side makes the theorem a direct consequence of the existence of a CSB generator.

5. Probabilistic Forecasting II

The evaluation of the quality of forecasting "on the average" (as in Section 2) is acceptable from the viewpoint of a big forecasting corporation.
dealing with many strings. However, often one is interested only in a particular string at hand. The rest of the paper is devoted to estimating the quality of probabilistic forecasting of individual strings.

To prove assertions about the quality of forecasting an individual string we should assume some connection between the measure \( P \) and the string \( x \). To this end we use a variant of Kolmogorov's notion of randomness (Martin-Löf, 1966, Sect. 2; Kolmogorov, 1983), in some way explicating the expression "the device generates the string \( x \)." First of all we recall some definitions. An algorithm \( \mathcal{A} \) is called a mode of description if it is \( \nu \)-invariant. Let \( x \in \mathbb{R}, i \in \emptyset \), and \( \mathcal{A} \) be a mode of description. The value

\[
K_{\mathcal{A}}(x \mid i) = \min \{ \ell(p) \mid p \in \mathbb{B}^*, \mathcal{A}(p) = x \}
\]

(as usual we set \( \min \emptyset = \infty \)) is called the \( \mathcal{A} \)-complexity of \( x \) given \( i \). There exists a mode of description \( \mathcal{S} \) for which the function \( \lambda x_i K_{\mathcal{S}}(x \mid i) \) is smallest to within an additive constant; due to Lemma 1 the standard reasoning from (Kolmogorov, 1965; Solomonoff, 1964) applies. Let us fix such \( \mathcal{S} \) and call the value \( K(x \mid i) \), defined to be \( K_{\mathcal{S}}(x \mid i) \), the Kolmogorov complexity of \( x \) given \( i \).

Chaitin (1975) and Levin (1974) independently introduced a variant of Kolmogorov complexity having some important technical advantages. It is defined as follows. A mode of description \( \mathcal{A} \) is called self-delimiting if for all \( i \) the set \( \{ p \in \mathbb{B}^* \mid \mathcal{A}(p) \text{ converges} \} \) is prefix-free (i.e., contains at most one prefix of any string). Among the self-delimiting modes of description there is a \( \mathcal{S}' \) for which the function \( K_{\mathcal{S}'}(x \mid i) \) is smallest to within an additive constant. We fix such \( \mathcal{S}' \), denote \( K_{\mathcal{S}'}(x \mid i) \) by \( H(x \mid i) \), and call this value the self-delimiting complexity of \( x \) given \( i \). It is easy to see (and well known) that, for some constant \( c \),

\[
K(x \mid i) - c \leq H(x \mid i) \leq K(x \mid i) + 2 \log K_{\mathcal{S}'}(x \mid i) + c, \quad \forall x, i.
\]

We assume that some \( \nu \)-invariant algorithm, when given \( x \in \mathbb{B}^n \) and using an oracle for an \( n \)-measure \( P \), for any \( n \), calculates \( P(x) \) with any given accuracy and, moreover, determines whether \( P(x) = 0 \). (This is a requirement on the chosen system of notation \( \nu \); see Section 1.)

Kolmogorov (1983) proposes to measure the degree of nonrandomness of a string \( x \in \mathbb{B}^n \) with respect to the uniform \( n \)-measure with \( n - K(x \mid n) \). Analogously, for an arbitrary \( n \)-measure \( P \) we put

\[
d_P(x) = - \log P(x) - H(x \mid P)
\]

(cf. (Levin, 1984, Sect. 1.1)); it is easy to see that \( d_P(x) \geq -c \) for all \( P \) and \( x \), where \( c \) is some constant. In the sequel we shall assume that \( d_P(x) \), the
deficiency of randomness of \( x \) with respect to \( P \), is small. Such an assumption is justified by the well known inequality

\[
P\{d_p(x) > \beta\} < 2^{-\beta},
\]

which holds for any integer \( \beta \). E.g., a person who believes that \( x \) is randomly selected from the measure \( P \) will usually believe that \( d_p(x) \leq 10 \), since the probability of violating this inequality is less than \( 10^{-3} \).

Let us prove (14):

\[
P\{d_p(x) > \beta\} = P\{P(x) < 2^{-H(x|P) - \beta}\} < \sum_x 2^{-H(x|P) - \beta} \leq 2^{-\beta}.
\]

Let \( x \in \{0,1\}^n \) be generated in accordance with a measure \( P \). For the evaluation of the quality of forecasting the string \( x \) by a forecasting system \( \mu \) we use in addition to \( q_\mu(x, P) \) the following two quantities:

\[
q_\mu(x, P) = \sum_{i=0}^{n-1} \rho(P(1|x^i), \mu(x^i))
\]

(the lower error of forecasting) and

\[
\bar{q}_\mu(x, P) = \sum_{i=0}^{n-1} \bar{\rho}(P(1|x^i), \mu(x^i))
\]

(the upper error of forecasting). The functions \( \rho \) and \( \bar{\rho} \) of the type \([0, 1]^2 \rightarrow \mathbb{R}\) are defined by the equalities

\[
\rho(u, v) = (\ln e)((u^{1/2} - v^{1/2})^2 + ((1 - u)^{1/2} - (1 - v)^{1/2})^2),
\]

\[
\bar{\rho}(u, v) = (\ln e)\frac{(u - v)^2}{v(1-v)},
\]

respectively (we set \( 0/0 = 0 \) and \( a/0 = \infty \) if \( a \neq 0 \)). If we ignore numeric factors, these are the distances of Hellinger and \( \chi^2 \) respectively between distributions \((u, 1-u)\) and \((v, 1-v)\). When \( u \) and \( v \) are close to each other (in some sense), \( \rho(u, v) \) is close to \( \frac{1}{2}\rho(u, v) \) and \( \bar{\rho}(u, v) \) is close to \( 2\rho(u, v) \) (Borovkov, 1984, p. 194).

In the next theorem quality of universal forecasting of individual strings is estimated. Regrettably, such estimates are harder to obtain and involve some additional error terms as compared with those "on the average."

**Theorem 8.** There exist an algorithm \( \Psi \), a constant \( c \), and a polynomial \( f \) such that the following holds. Let \( n, \Gamma \in \mathbb{Z}^+ \). For \( \Gamma \geq c \) all the values

\[
\pi(y) = \Psi(n, \Gamma, y), \quad y \in \{0,1\}^n.
\]
are computed within time $T$ and $\pi$ is an $n$-forecasting system. If $\alpha, \gamma \in \mathbb{Z}^+$ satisfy the inequality $f(n^{2\gamma}) \leq T$, $P$ is an $n$-measure, $x \in \mathbb{B}^n$, and $\mu$ is an $(\alpha, \gamma)$-simple $n$-forecasting system, then

$$q_\pi(x, P) \leq \tilde{q}_\mu(x, P) + \alpha + 1.01 \log \alpha + 3d_\mu(x) + 3H(\mu | P) + 3H(\Gamma | n) + c. \quad (15)$$

Inequality (15) may be replaced with a clearer one,

$$q_\pi(x, P) \leq \tilde{q}_\mu(x, P) + 4.01\alpha + 3d_\mu(x) + 3.01K(\Gamma | n) + c.$$

Let us convince ourselves that $\Psi$ is in some sense "universal." For a forecasting system $\mu$ used in practice, $\alpha$ and $H(\mu | P)$ are usually small constants and $\gamma$ is not too large. It already was mentioned that we are concerned with the case of small $d_\mu(x)$. The value $T$ is an estimate of the computational resources of the forecaster, so the (somewhat mysterious) term $H(\Gamma | n)$ is natural to assume small. So inequality (15) shows that the error $q_\pi(x, P)$ of forecasting by $\Psi$ is not much worse than the error $\tilde{q}_\mu(x, P)$ of forecasting by $\mu$.

Now we consider the question of optimality of the algorithm $\Psi$ and of the bounds from Theorem 8. Let us investigate the optimality of Theorem 8 in respect to the addends $\tilde{q}_\mu(x, P)$ and $\alpha$ (the latter is important for considerations such as those in Sect. 7) on the right-hand side of (15).

**Theorem 9.** There exist such a constant $c$ and a polynomial $f$ that the following holds. Let $n \in \mathbb{Z}^+$, $\pi$ be an $n$-forecasting system, and $\eta, \alpha \in \mathbb{Z}^+$ be such that $\eta + \alpha < n$. One can find an $(\alpha + c, f(n))$-simple $n$-forecasting system $\mu$, an $n$-measure $P$, and a string $x \in \mathbb{B}^n$ such that $d_\mu(x) \leq 0$, $H(\mu | P) \leq c$, $\tilde{q}_\mu(x, P) \leq (\log e)(\eta + 1)$, and

$$q_\pi(x, P) \geq (\log e)(2 - 2^{1/2})(\eta + \alpha).$$

This theorem shows that, as concerns $\tilde{q}_\mu(x, P)$ and $\alpha$, estimate (15) is sharp to within a constant factor. The next theorem concerns the running time of $\Psi$.

**Theorem 10.** Suppose one-way functions exist. Let $\delta > 0$ be arbitrarily small. There is such a polynomial $f$ that the following holds. Let an algorithm $Q$ be such that the values

$$\pi_n(y) = Q(n, y), \quad n \in \mathbb{Z}^+, \ y \in \mathbb{B}^{<n},$$

are computed within polynomial time and the functions $\pi_n$ so defined are $n$-forecasting systems. Then for infinitely many $n$ there are such an $(n^\delta, f(n))$-simple $n$-measure $P$ and a string $x \in \mathbb{B}^n$ that $d_\mu(x) \leq 0$ and

$$q_\pi(x, P) \geq (\log e)(2 - 2^{1/2} - \delta)n. \quad (16)$$
If we set \( r_n(y) = \frac{1}{2}, \forall y \in \mathbb{B}^{|n|} \), then
\[
q_{r_n}(x, P) \leq n q(0, \frac{1}{2}) = (\log e)(2 - 2^{1/2})n.
\]
So, \( \mathcal{Q} \) is a poor forecaster as concerns \( P \) and \( x \).

6. Probabilistic Forecasting III

In the previous section the quality of the universal forecasting algorithms was measured in a more liberal way than the quality of rivals. In this section both the qualities will be measured uniformly, viz., by Kullback-Leibler distance. The price we must pay is the restriction
\[
\varepsilon \leq P(1 | y) \leq 1 - \varepsilon, \quad \forall y \in \mathbb{B}^{|n|},
\]
where \( \varepsilon > 0 \), imposed on the "true" measure; we assume \( \varepsilon \in \mathbb{D} \). \( n \)-measures satisfying (17) will be called \((n, \varepsilon)\)-measures. We suppose that the binary expansion of \( \varepsilon \) is known to all algorithms. To avoid unnecessary complications we also suppose that \( \varepsilon \geq 2^{-c}, \) \( c \) being a constant, throughout (so the binary expansion of \( \varepsilon \) is of polynomial length). The reader is advised to treat \( \varepsilon \) as a constant: when \( \varepsilon \to 0 \), our estimates of the quality of forecasting deteriorate.

We say that \( \mu \) is an \((n, \varepsilon)\)-forecasting system if it is an \( n \)-forecasting system such that \( \varepsilon \leq \mu(y) \leq 1 - \varepsilon, \ y \in \mathbb{B}^{|n|} \). When the "true measure" is known to be an \((n, \varepsilon)\)-measure, it is natural to consider only \((n, \varepsilon)\)-forecasting systems, in particular to replace Laplace's rule of succession (1) with the forecasting system
\[
\lambda(y) = \varepsilon \lor \frac{k + 1}{h(y) + 2} \land (1 - \varepsilon), \quad y \in \mathbb{B}^{|n|},
\]
where \( k \) is the number of 1s in \( y \). Now we formulate analogs of Theorems 8 through 10.

**Theorem 11.** There exist an algorithm \( \mathcal{B} \), a constant \( c \) and a polynomial \( f \) such that the following holds. Let \( \delta > 0 \) be arbitrarily small, \( n, \Gamma \in \mathbb{Z}^\times \), \( \Gamma \geq c \), and \( \varepsilon \in \mathbb{D} \). All the values
\[
\pi(y) = \mathcal{B}(n, \varepsilon, \Gamma, y), \quad y \in \mathbb{B}^{|n|},
\]
are computed within time \( \Gamma \). The functions \( \pi \) are \( n \)-forecasting systems. If \( \alpha, \gamma \in \mathbb{Z}^\oplus \) satisfy \( f(n 2^\gamma y) \leq \Gamma, \) \( P \) is an \((n, \varepsilon)\)-measure, \( x \in \mathbb{B}^n \), and \( \mu \) is an \((\alpha, \gamma)\)-simple \( n \)-forecasting system, then
\[
q_\alpha(x, P) \leq (q + \alpha) + c' \varepsilon^{-\delta}((q + \alpha) \log \log(q + \alpha))^{1/2}
+ c' \varepsilon^{-\delta}(A(q + \alpha))^{1/2} + c' \varepsilon^{-\delta}A,
\]
where \( q = q_\mu(x, P) \), \( A = d_P^\oplus(x) + K(\mu \mid P) + K(\Gamma \mid n) \), and \( c' \) depends only on \( \delta \).

Inequality (19) implies
\[
q_n(x, P) \leq (1 + \delta) q_\mu(x, P) + c''(\alpha + d_\rho(x) + K(\Gamma \mid n)),
\]
where \( c'' \) may depend on \( \delta \) and \( \varepsilon \).

**Note 4.** The last inequality shows that with probability close to 1 the error \( q_n(x, P) \) is not much bigger than the error \( q_\mu(x, P) \) (\( \alpha \) and \( K(\Gamma \mid n) \) are supposed small). Theorem 1, which asserts that the mean value \( A_\mu(P) \) of \( q_\mu(x, P) \) is not much bigger than the mean value \( A_\mu(P) \) of \( q_\mu(x, P) \), implies nothing like this. Indeed, for arbitrarily small \( \varepsilon > 0 \) there are positive random variables \( \xi_1 \) and \( \xi_2 \) such that \( E\xi_1 \leq \varepsilon E\xi_2 \) and \( P\{\xi_2 < \varepsilon \xi_1\} > 1 - \varepsilon \).

Let us introduce the notation
\[
\mathcal{H}(\varepsilon) = -\varepsilon \log \varepsilon - (1 - \varepsilon) \log (1 - \varepsilon).
\]
Recall that the function \( \mathcal{H}(\varepsilon) \), entropy of a distribution \( (\varepsilon, 1 - \varepsilon) \), satisfies \( \lim_{\varepsilon \to 0} \mathcal{H}(\varepsilon) = 0 \).

**Theorem 12.** Some constant \( c \) and polynomial \( f \) satisfy the following. Let \( n \in \mathbb{Z}^\oplus, \pi \) be an \( n \)-forecasting system, \( \varepsilon \in \mathbb{D}, \beta \in \mathbb{Z}^\oplus, \) and \( \eta, \alpha \in \mathbb{Z}^\oplus \) do not exceed \( n/10 \). There are an \((\alpha + c, f(n))\)-simple \((n, \varepsilon)\)-forecasting system \( \mu \), an \((n, c)\)-measure \( P \), and a string \( x \in B^n \) such that \( d_\rho(x) \leq \beta + 3, K(\mu \mid P) \leq c, q_\mu(x, P) \leq \eta + 1, \) and
\[
q_n(x, P) \geq (1 + 2^{-\beta} \log \varepsilon)(\eta + (1 - \mathcal{H}(\varepsilon))\alpha).
\]

This theorem shows that, in the case where \( \varepsilon \) is a small constant, \( 1 \ll 2^{d_\rho(x)} \), and \( A \ll q + \alpha \) (\( u \ll v \) means that the ratio \( u/v \) is small), inequality (19) is nearly sharp (the optimality with respect to \( \alpha \) is needed for applications described in the next section). When \( \varepsilon \) is not treated as a constant, (19) is nearly sharp provided that
\[
\varepsilon \ll 1, \quad |\log \varepsilon| \ll 2^{d_\rho(x)}, \quad \frac{A + \log \oplus \log(q + \alpha)}{q + \alpha} \ll e^{2\delta}.
\]

**Theorem 13.** Suppose one-way functions exist. Let \( \delta > 0 \) be arbitrarily small. There are such a constant \( c \) and a polynomial \( f \) that the following holds. Let an algorithm \( \mathcal{Q} \) compute the values
\[
\pi_{n, \varepsilon}(y) = \mathcal{Q}(n, \varepsilon, y), \quad n \in \mathbb{Z}^\oplus, \varepsilon \in \mathbb{D}, y \in B^{<n},
\]
within polynomial time and the functions $\pi_{n, \varepsilon}$ be n-forecasting systems. Then for infinitely many $n$ and for all $\varepsilon \in \mathbb{D}$, $\beta \in \mathbb{Z}^\oplus$ there exist an $(n^\beta, f(n))$-simple $(n, \varepsilon)$-measure $P$ and a string $x \in \mathbb{B}^n$ such that $d_P(x) \leq \beta + 1$ and

$$q_{\pi_{n, \varepsilon}}(x, P) \geq (1 - H(\varepsilon)) + 2^{-\beta} \log \varepsilon n. \quad (20)$$

The poor performance of $\Omega$ (at least provided that $|\log \varepsilon| \leq 2^\beta$) is demonstrated as before: if $\pi_{n, \varepsilon}(y) = \frac{1}{2}$ for all $y \in \mathbb{B}^{-n}$,

$$q_{\pi_{n, \varepsilon}}(x, P) \leq n\log(e, \frac{1}{2}) = (1 - H(\varepsilon))n.$$

### 7. Accuracy of Forecasting

Even if we knew that we got the best forecasts attainable at our computer, there would remain an important question about the size of the error of forecasting. Theorems 1, 8, and 11 sometimes answer this question too. We consider here only forecasting individual strings treated in Theorem 11.

Let $\mathcal{P}$ be a set of $n$-measures and $\Pi$ a finite set of $n$-forecasting systems. We call $\Pi$ a $q$-net for $\mathcal{P}$ if for any $P \in \mathcal{P}$ there exists $\pi \in \Pi$ such that $q_\pi(x, P) < 1$ for any $x \in \mathbb{B}^n$. Let $A$ be the minimum of the cardinalities of $q$-nets for $\mathcal{P}$. Then $\log A$ is called the $q$-entropy of $\mathcal{P}$. Usually in practice for a set $\mathcal{P}$ of $n$-measures whose entropy does not exceed $A$ an algorithm $\mathfrak{A}$ can be readily pointed out such that $\mathcal{P}$ admits a net consisting of $(\alpha + c, \gamma(n))$-simple, with respect to $\mathfrak{A}$, $n$-forecasting systems, where $c$ is a small constant and $\gamma$ is a slowly growing polynomial. Let $\mathfrak{B}$ be the algorithm from Theorem 11.

**Corollary 1.** Let $\varepsilon \in \mathbb{D}$. An integer $c$ and a polynomial $f$ exist satisfying the following. If for some $n, \alpha, \gamma \in \mathbb{Z}^\oplus$, a family $\mathcal{P}$ of $(n, \varepsilon)$-measures has a $q$-net consisting of $(\alpha, \gamma)$-simple forecasting systems, then for all $P \in \mathcal{P}$, $x \in \mathbb{B}^n$ and integers $\Gamma \geq f(n^{2\gamma})$, the $n$-forecasting system $\pi$ defined by $\pi(y) = \mathfrak{B}(n, \varepsilon, \Gamma, y)$ satisfies

$$q_\pi(x, P) \leq \alpha + c(\alpha \log \log \alpha)^{1/2} + c(A\alpha)^{1/2} + cA,$$

where $A = d_P^\oplus(x) + K(\Gamma | n)$.

**Proof.** Note that given $P$ a forecasting system $\mu$ can be effectively found such that $q_\mu(x, P) < 1$. Applying Theorem 11 completes the proof.  

E.g., if $\varepsilon > 0$ and $\mathcal{P}$ is the set $\{B_\theta \mid \theta \in [\varepsilon, 1 - \varepsilon]\}$ of Bernoulli $n$-measures with probability $\theta$ of success, one can take the set

$$\left\{ \mu \mid \mu \equiv \frac{a\delta}{n^{1/2}} \in [\varepsilon, 1 - \varepsilon], \ a \in \mathbb{Z} \right\},$$
264

VLADIMIR G. VOVK

$\delta > 0$ being some rational constant, as a $q$-net. Corollary 1 implies that the error of forecasting by $\mathcal{B}$ asymptotically does not exceed $\frac{1}{2}\log n$. It may be easily verified that if $n$ is large, $P$ is a Bernoulli $n$-measure, and $x \in \mathcal{B}^n$ satisfies $d_r(x) \leq c$, then $q_j(x, P)$ asymptotically coincides with $\frac{1}{2}\log n$ ( $\lambda$ is the version of Laplace's rule of succession given by (18)). In (Vovk, 1989) other important statistical models are considered as well (however, in a somewhat different framework).

8. BOUNDS ON DEFICIENCY OF RANDOMNESS

Let $d_r(x|i)$ stand for $-\log P(x) - H(x|P, i)$. We first sketch the idea of the proof of Theorem 8. Until Lemma 4 we disregard conditions (i.e., variables standing after the vertical line in expressions for deficiency of randomness and self-delimiting complexity) and designate with "=" and "\(\leq\)" approximate equality and inequality respectively. It is Lemma 4 which plays a key role. Let a string $x$ be generated in accordance with a measure $P$, i.e., let $d_r(x)$ be small, and $Q$ be another measure. That lemma shows that $d_r(x)$ lies approximately between the errors $q(x, P)$ and $q(x, P)$ of forecasting the string $x$ by $Q$. "Universal forecasting systems" $\pi$ were constructed in Lemma 3; $\pi$ is the forecasting system corresponding to a weighted mean of the measures corresponding to simple forecasting systems, and any $(x, \gamma)$-simple forecasting system involved in forming this mean is taken with the weight $\geq 2^{-a}$. It follows from the definition of $d$ that if the forecasting system $\mu$ is $(x, \gamma)$-simple, then

$$d_\pi(x) = -\log P(x) - H(x) \leq -\log \mu(x) + \alpha - H(x) = d_\mu(x) + \alpha.$$ 

In accordance with Lemma 4 we have

$$q(x, P) \leq d_\pi(x) \leq d_\mu(x) + \alpha \leq q_\mu(x, P) + \alpha.$$ 

Thus the quality of forecasting by $\pi$ (measured with $q$) turns out to be not much lower than that by $\mu$ (measured with $q$).

Theorem 11 is proved in a similar way; the role of Lemma 4 is played by Lemma 6 which asserts that $q(x, P)$ approximately equals $d_r(x)$.

**Lem 4.** For some constant $c$ the following holds. Let $n \in \mathbb{Z}^+$, $x \in \mathcal{B}^n$, and $P, Q$ be $n$-measures. Then

$$\sum_{i=0}^{n-1} \rho(P(1|x^i), Q(1|x^i)) - d_r(x|Q) - c \leq d_\rho(x|P) \leq \sum_{i=0}^{n-1} \bar{\rho}(P(1|x^i), Q(1|x^i)) + 2d_r(x|Q) + c.$$
Note 5. The lower estimate of Lemma 4 shows that any two simple and "successful" forecasting systems agree. This phenomenon is known from (Dawid, 1985) (with "computable" for "simple" and "computably calibrated" for "successful").

To prove Lemma 4 we need one more lemma.

**Lemma 5.** Let

\[ \{ R_{n,\kappa, i} \mid n \in \mathbb{Z}^+, \kappa \in \mathcal{F}, i \in \mathcal{C} \}, \]

where \( R_{n,\kappa, i} \) is an \( n \)-measure for all \( n, \kappa, \) and \( i \), be a computable family. There exists a constant \( c \) such that, for all \( n, \kappa, i, \) and \( x \in \mathbb{B}^n \),

\[ \log R_{n,\kappa, i}(x) \leq -H(x|n, i) + H(\kappa) + c. \]

**Proof.** Note that there are constants \( c_1 \) and \( c_2 \) satisfying

\[ H(x|n, i, \kappa) \leq -\log R_{n,\kappa, i}(x) + c_1, \]

\[ H(x|n, i) \leq H(x|n, i, \kappa) + H(\kappa) + c_2. \]

**Proof of Lemma 4. The Lower Bound.** Let \(-1 < \kappa < 0\) be a rational constant. Define an \( n \)-measure \( R = R_{n,\kappa, p, Q} \) by

\[ R(j|y) = (P^{1+\kappa}(j|y) Q^{-\kappa}(j|y)) \left( \sum_{j=0}^{1} P^{1+\kappa}(j|y) Q^{-\kappa}(j|y) \right), \]

\[ y \in \mathbb{B}^{<n}, \quad j \in \mathbb{B} \]

(21)

(degenerate cases will be considered separately). Due to Lemma 5 and the definition of \( d \),

\[ \log R(x) \leq \log P(x) + d_P(x|Q) + H(\kappa) + c_1, \quad \forall x \in \mathbb{B}^n \]

\((c_1, c_2, \ldots \) are constants). Using the equalities

\[ R(x) = \prod_{i=0}^{n-1} R(x_i|x'), \quad P(x) = \prod_{i=0}^{n-1} P(x_i|x'), \]

and rewriting \( R(x_i|x') \) in accordance with (21), we obtain

\[ \log(P^{\kappa}(x)/Q^{\kappa}(x)) \leq \log \left( \prod_{i=0}^{n-1} \left( \sum_{j=0}^{1} P^{1+\kappa}(j|x') Q^{-\kappa}(j|x') \right) \right) + d_P(x|Q) + H(\kappa) + c_1. \]
Since
\[ |\log(P(x)/Q(x)) - (d\rho(x|P) - d\rho(x|Q))| \leq c_2, \]
it follows that
\[
d\rho(x|P) - d\rho(x|Q) \geq \kappa^{-1} \sum_{i=0}^{n-1} \log \left( \sum_{j=0}^{i} P(j|x^i) \times (P(j|x^i)/Q(j|x^i))^\kappa \right)
\]
\[ + (d\rho(x|Q) + H(\kappa) + c_3)/\kappa. \tag{22} \]

We set \( \kappa \) to \(-\frac{1}{2}\). The assertion that we are proving can now be deduced from
\[
2 \log \left( \sum_{i=0}^{1} P(j|x^i)(P(j|x^i)/Q(j|x^i))^{1/2} \right)
\]
\[ = (2 \log e) \log \left( \sum_{j=0}^{1} (P(j|x^i) Q(j|x^i))^{1/2} \right)
\]
\[ = (2 \log e) \log \left( 1 - \frac{1}{2} \sum_{j=0}^{1} ((P(j|x^i))^{1/2} - (Q(j|x^i))^{1/2})^2 \right)
\]
\[ \leq -\rho(P(1|x^i), Q(1|x^i)). \]

Let us now consider the degenerate cases. When \( P(x) = 0 \) or \( Q(x) = 0 \) we have \( d\rho(x) = \infty \) or \( d\rho(x) = \infty \), so the desired inequality trivially holds. It remains to consider the case that the denominator in (21) vanishes. But in this case we again obtain \( P(x) = 0 \) or \( Q(x) = 0 \).

The Upper Bound. This time we set \( \kappa = 1 \) in (21). The same calculation as before yields (22) with \( \leq \) replaced by \( \geq \) (since \( \kappa \) is now positive). Note that
\[
\log \sum_{j=0}^{1} \frac{P^2(j|x^i)}{Q(j|x^i)} = (\log e) \log \left[ 1 + \sum_{j=0}^{1} \frac{(P(j|x^i) - Q(j|x^i))^2}{Q(j|x^i)} \right]
\]
\[ \leq \rho(P(1|x^i), Q(1|x^i)). \]

It remains to consider the degenerate cases. When \( P(x) = 0 \) or \( P(j|x^i) \neq Q(j|x^i) = 0 \) for some \( j \) and \( i \), the desired inequality trivially holds (its right-hand part is infinite), so we suppose that \( P(x) \neq 0 \) and \( Q(j|x^i) = 0 \) implies \( P(j|x^i) = 0 \). For \( P(j|x^i) = Q(j|x^i) = 0 \) (in this case necessarily \( j = 1 - x_i \)) we put \( P^2(j|x^i)/Q(j|x^i) = 0 \) in (21) and act as before.
**Lemma 6.** Let a constant $\delta > 0$ be arbitrarily small. For some integer $c$ the following holds. Let $n \in \mathbb{Z}^+$, $\varepsilon \in \mathbb{D}$, $x \in \mathbb{R}^n$, $P$ be an $n$-measure, and $Q$ be an $(n, \varepsilon)$-measure. Then

$$|d - E| \leq c e^{-\delta} (E^\oplus \log \log E^\oplus)^{1/2} + c e^{-\delta} (E^\oplus d^\oplus_P(x \mid Q))^{1/2} + c d_P^\oplus(x \mid Q),$$

$$E - d \leq c e^{-\delta} (d^\oplus \log \log d^\oplus)^{1/2} + c e^{-\delta} (d^\oplus d^\oplus_P(x \mid Q))^{1/2} + c e^{-\delta} d_P^\oplus(x \mid Q),$$

where

$$d = d_Q(x \mid P), \quad E = \sum_{i=0}^{n-1} \rho(P(1 \mid x^i), Q(1 \mid x^i)).$$

First we prove an auxiliary assertion.

**Lemma 7.** Let functions $f_1$ and $f_2$ be defined by

$$f_k(p, q) = p \log k(p/q) + (1 - p) \log k((1 - p)/(1 - q)).$$

There is a constant $c$ such that, for all $\varepsilon \in \mathbb{D}$,

$$f_2(p, q) \leq (-c \log \varepsilon) f_1(p, q),$$

(24)

where $p$ ranges over $[0, 1]$ and $q$ over $[\varepsilon, 1 - \varepsilon]$.

**Proof.** By the symmetry without loss of generality we may suppose $q \leq \frac{1}{2}$. We also suppose $p \neq q$ and replace $\log$ by $\ln$. It is easy to check that

$$f_1 \sim p \ln(p/q), \quad f_2 \sim p \ln^2(p/q), \quad f_2/f_1 \sim \ln(p/q)$$

(we omit the arguments $p, q$ of both the functions) as $p \to 0$, $q \to 0$, $p/q \to \infty$, and

$$f_1 \sim q, \quad f_2 \sim p \ln^2(p/q) + q^2, \quad f_2/f_1 \sim (p/q) \ln^2(p/q) + q \to 0$$

as $p \to 0$, $q \to 0$, $p/q \to 0$. Therefore, in some square $0 \leq p < \delta_1$, $0 \leq q < \delta_1$ we have $f_2/f_1 < 1 + 2 \ln(p/q)$ except for $p, q$ satisfying $1/c_1 < p/q < c_1$, where $c_1$ is some constant. Let $\approx$ stand for coincidence to within a constant factor. Then $p/q$ and $(1 - p)/(1 - q)$ lie in $]1/c_1, c_1[$.

When $p/q$ and $(1 - p)/(1 - q)$ lie in $]1/c_1, c_1[$,

$$f_1 = -p \ln \frac{q}{p} - (1 - p) \ln \frac{1 - q}{1 - p}$$

$$= -p \left( \frac{q}{p} - 1 - \frac{1}{2} \left( \left( \frac{q}{p} - 1 \right) / \xi_1 \right)^2 \right)$$

$$- (1 - p) \left( \frac{1 - q}{1 - p} - 1 - \frac{1}{2} \left( \left( \frac{1 - q}{1 - p} - 1 \right) / \xi_2 \right)^2 \right)$$

$$\approx p \left( \frac{q - 1}{p} \right)^2 + (1 - p) \left( \frac{1 - q - 1}{1 - p} \right)^2 = \frac{(p - q)^2}{p(1 - p)},$$

$$f_2 = -q \ln \frac{q}{p} - (1 - q) \ln \frac{1 - q}{1 - p}$$

$$= -q \left( \frac{q}{p} - 1 - \frac{1}{2} \left( \left( \frac{q}{p} - 1 \right) / \xi_1 \right)^2 \right)$$

$$- (1 - q) \left( \frac{1 - q}{1 - p} - 1 - \frac{1}{2} \left( \left( \frac{1 - q}{1 - p} - 1 \right) / \xi_2 \right)^2 \right)$$

$$\approx (1 - q) \left( \frac{q - 1}{p} \right)^2 + (1 - p) \left( \frac{1 - q - 1}{1 - p} \right)^2 = \frac{(q - p)^2}{q(1 - q)}.$$
where $\xi_1$, $\xi_2$, $\xi'_1$, $\xi'_2$ lie between 1 and $q/p$, $p/q$, $(1-q)/(1-p)$, $(1-p)/(1-q)$, respectively (we used Lagrange's form of the remainder in Taylor's formula); hence $f_1 \approx f_2$. Suppose that $p > 6$, and $q$ approaches 0. Then

$$f_1 \sim p \ln(p/q), \quad f_2 \sim p \ln^2(p/q), \quad f_2/f_1 \sim \ln(p/q).$$

So, in some rectangle $\delta_1 \leq p \leq 1$, $0 \leq q < \delta_2$ with $\delta_2 < \delta_1$ we have $f_2/f_1 < 2 \ln(p/q)$. The complement $A$ of all the considered regions is closed, and the function $f_2/f_1$ is defined and continuous on $A$. So, $f_2/f_1$ is bounded from above on $A$. We have shown that

$$f_2 \leq (c_2 \vee 2 \ln(p/q)) f_1$$

for some constant $c_2$, whence immediately follows (24).

Proof of Lemma 6. The 1st Inequality. First we prove

$$d_Q(x \mid P) - E \leq c e^{-\delta} (E \oplus \lb \oplus E^{\oplus})^{1/2} + c e^{-\delta} (E \oplus d_P^{\oplus}(x \mid Q))^{1/2} + c d_P^{\oplus}(x \mid Q). \quad (25)$$

We define an $n$-measure $R$ by (21) ($\kappa > 0$ will be chosen in the sequel). As before we get the inequality opposite to (22) ($\kappa$ is positive). By virtue of the inequalities

$$\lb(1 + t) = \ln(1 + t) \lb e \leq t \lb e,$$

$$2^t = e^{t \ln 2} \leq 1 + t \ln 2 + (t^2/2) \ln 2 \leq e^{t \ln 2}$$

we obtain

$$d_Q(x \mid P) - d_P(x \mid Q)$$

$$\leq \kappa^{-1} \sum_{i=0}^{n-1} \lb \left[ \sum_{j=0}^{1} P(j \mid x') 2^{\kappa \lb(P(j \mid x') \mid Q(j \mid x'))} \right]$$

$$+ (d_P(x \mid Q) + H(\kappa) + c_3)/\kappa$$
\[
\sum_{j=0}^{n-1} \left[ \sum_{j'=0}^{1} \frac{P(j \mid x') \log(P(j \mid x')/Q(j \mid x'))}{\log_2(P(j \mid x')/Q(j \mid x'))} g_1(\varepsilon) \right] + (d_\rho(x \mid Q) + H(\kappa) + c_3)/\kappa;
\]

here \( g_1(\varepsilon) \) is the least upper bound of numbers

\[
\left( \frac{1}{2} \right) \log_2 \left( \frac{P(j \mid x')}{Q(j \mid x')} \right) \leq \varepsilon - |\kappa|.
\]

Taking into account Lemma 7 we get

\[
d_\rho(x \mid P) - d_\rho(x \mid Q) \leq E + \kappa g_2(\varepsilon) E + (d_\rho(x \mid Q) + H(\kappa) + c_3)/\kappa,
\]

where \( g_2(\varepsilon) = (-\log_2 \varepsilon) g_1(\varepsilon) c_4 \). Take

\[
\kappa = 2^{\log_2 \kappa^*}, \quad \kappa^* = \delta' \wedge \left( (\log_2 E^\oplus + d^*)/E^* \right)^{1/2},
\]

\[
d^* = 2^{\log_2 d^+(x \mid Q)}, \quad E^* = 2^{\log_2 E^\oplus},
\]

where \( \delta' < \delta \), \( \delta' \in \mathbb{D} \). Note that \( g_2(\varepsilon) \leq c_5 \varepsilon^{-\delta} \). Let us consider two cases.

(i) \( \kappa^* < \delta' \). In this case \( \kappa \) as a function of \( P, Q, \) and \( x \) to within a constant factor coincides with

\[
\frac{1}{(\log_2 E^\oplus + d^+(x \mid Q))^{1/2}.}
\]

From (26) we get

\[
d_\rho(x \mid P) - d_\rho(x \mid Q) \leq E + c_6 g_2(\varepsilon)(E^\oplus \log_2 \log_2 E^\oplus + d^+(x \mid Q))^{1/2}
\]

\[
+ c_6 (E^\oplus d^+(x \mid Q))^{1/2} + c_6 H(\kappa)/\kappa + c_6 (E^\oplus)^{1/2}.
\]

As \( \kappa \) can be effectively found given \( d^* \) and \( E^* \),

\[
H(\kappa) \leq H(\log_2 E^\oplus) + H(\log_2 d^+(x \mid Q)) + c_7
\]

\[
\leq c_8 (\log_2 \log_2 E^\oplus + \log_2 \log_2 d^+(x \mid Q));
\]

the last two inequalities imply (25).

(ii) \( \kappa^* = \delta' \). In this case \( d^+(x \mid Q) \geq E^\oplus/c_9 \), where \( c_9 \) may depend on \( \delta \). By the direct substitution of \( 2^{\log_2 \delta} = \delta' \) for \( \kappa \) in (26) one can easily verify that (25) is satisfied.

So (25) is proved. It remains to prove

\[
d_\rho(x \mid P) - E \geq -c\varepsilon^{-\delta}(E^\oplus \log_2 \log_2 E^\oplus)^{1/2}
\]

\[
- c\varepsilon^{-\delta}(E^\oplus d^+(x \mid Q))^{1/2} - cd^+(x \mid Q).
\]
The proof is analogous with that of (25), except that \( \kappa \) must be taken with the reversed sign:

\[
\kappa = -2^{1}b^{*}k^{*}l.
\]

**The 2nd Inequality.** Let us denote

\[
d_{p} = d_{p}(x \mid Q); \quad d_{Q} = d_{Q}(x \mid P).
\]

Due to the 1st inequality we get

\[
E - ce^{-\delta}(F^{\oplus} lb^{\oplus} lb^{\oplus} F^{\oplus})^{1/2} - ce^{-\delta}(F^{\oplus} d_{p}^{\oplus})^{1/2} \leq d_{Q} + cd_{p}^{\oplus}.
\] (27)

We consider two cases:

(i) \( d_{p}^{\oplus} \leq lb^{\oplus} lb^{\oplus} E^{\oplus} \). In this case (27) implies

\[
E - 2ce^{-\delta}(E^{\oplus} lb^{\oplus} lb^{\oplus} E^{\oplus})^{1/2} \leq d_{Q} + c lb^{\oplus} lb^{\oplus} E^{\oplus}.
\] (28)

Hence

\[
E \leq 2d_{Q} + g(\varepsilon),
\] (29)

where

\[
g(\varepsilon) = \sup_{E} (-E + 4ce^{-\delta}(E^{\oplus} lb^{\oplus} lb^{\oplus} E^{\oplus})^{1/2} + 2c lb^{\oplus} lb^{\oplus} E^{\oplus})
\leq \sup_{E} (-E + (c_{1} e^{-\delta} + c_{2}) E^{3/4} + c_{3})
= c_{4}(c_{1} e^{-\delta} + c_{2})^{4} + c_{3} \leq c_{5} e^{-4\delta}
\]

\((c_{1}, c_{2}, \ldots \) are integers depending only on \( \delta \)). From (27), (29) we obtain

\[
E \leq d_{Q} + c_{6} e^{-\delta}((d_{Q}^{\oplus} + g(\varepsilon)) lb^{\oplus} lb^{\oplus} (d_{Q}^{\oplus} + g(\varepsilon)))^{1/2} + c_{6} d_{p}^{\oplus}
\leq d_{Q} + c_{7} e^{-5\delta}(d_{Q}^{\oplus} lb^{\oplus} lb^{\oplus} d_{Q}^{\oplus})^{1/2} + c_{7} d_{p}^{\oplus},
\]

which agrees with (23).

(ii) \( lb^{\oplus} lb^{\oplus} E^{\oplus} \leq d_{p}^{\oplus} \). Now (27) implies

\[
E^{\oplus} - 2ce^{-\delta}(E^{\oplus} d_{p}^{\oplus})^{1/2} \leq d_{Q} + c_{8} d_{p}^{\oplus}.
\]

It is a square inequality with respect to \((E^{\oplus})^{1/2}\). Let us write down its solution as

\[
A \leq (E^{\oplus})^{1/2} \leq B.
\]
Since $A < 0$, it is equivalent to $E^\oplus \leq B^2$, i.e.,
\[
E^\oplus \leq d_Q + c_9 e^{-A(d_Q^\oplus d_P^\oplus)^{1/\gamma}} + c_{10} e^{-\gamma d_P^\oplus},
\]
which also agrees with (23).

Some versions of the inequalities from Lemmas 4 and 6 are proved in (Vovk, 1987, Theorem 1) and (Vovk, 1989, Theorem 2.2).

9. PROOFS OF THEOREMS 8, 9, 10

The "universal forecasting algorithm" $\mathfrak{B}$ was constructed in Lemma 3. We shall often use the "relativized" variants of Lemmas 4 and 6, e.g., the inequality
\[
d_Q(x \mid \Gamma, P) \leq \sum_{i=0}^{n-1} \tilde{\rho}(P(1 \mid x^i), Q(1 \mid x^i)) + 2d_P(x \mid \Gamma, Q) + c.
\]

Proof of Theorem 8. From Lemma 4 (the upper bound) it follows that
\[
d_\mu(x \mid \Gamma, P) \leq \tilde{q}_\mu(x, P) + 2d_P(x \mid \Gamma, \mu) + c_1
\]
(\(c_1, c_2, \ldots\) are constants). By the definition of $\mathfrak{B}$ (see Lemma 3),
\[
d_\mu(x \mid \Gamma, \mu, P) = -\log \pi'(x) - H(x \mid \pi', \Gamma, \mu, P)
\leq -\log \mu'(x) + \alpha + 1.01 \log \alpha + c_2 - H(x \mid \pi, \Gamma, \mu, P)
\leq -\log \mu'(x) - H(x \mid \mu', \Gamma, P) + \alpha + 1.01 \log \alpha + c_3
= d_\mu(x \mid \Gamma, P) + \alpha + 1.01 \log \alpha + c_3.
\]
To justify the last "\(\leq\)" note that \(\pi\) can be effectively found given \(\Gamma\) and \(n\) (and \(n\) can be extracted from \(P\) or \(\mu\)). Thus from (30) it follows that
\[
d_\pi(x \mid \Gamma; \mu, P) \leq \tilde{q}_\mu(x, P) + 2d_P(x \mid \Gamma, \mu) + \alpha + 1.01 \log \alpha + c_4.
\]
Using Lemma 4 (the lower bound) yields
\[
q_\pi(x, P) \leq \tilde{q}_\mu(x, P) + 2d_P(x \mid \Gamma, \mu) + \alpha + 1.01 \log \alpha + d_P(x \mid \Gamma, \mu, \pi) + c_5
\leq \tilde{q}_\mu(x, P) + 3d_P(x \mid \Gamma, \mu) + \alpha + 1.01 \log \alpha + c_6
\]
and, therefore, (15).

Proof of Theorem 9. Let \(m = \eta + \alpha\) and \(z \in \mathbb{B}^n\) be a string such that, for all \(i = 0, 1, \ldots, m - 1\),
\[
\pi(z^i) > \frac{1}{2} \Rightarrow z_i = 0,
\]
\[
\pi(z^i) < \frac{1}{2} \Rightarrow z_i = 1.
\]
The $P$ and $\mu$ are determined by the equalities
\[
P(z_i | z^i) = 1, \quad i = 0, 1, ..., m - 1,
\]
\[
P(1 | y^i) = \frac{1}{2}, \quad i = m, ..., n - 2,
\]
\[
\mu(y^i) = z_i, \quad i = 0, 1, ..., x - 1,
\]
\[
\mu(y^i) = \frac{1}{2}, \quad i = x, ..., n - 1,
\]
\[
P(1 | y^{n-1}) = x/n
\]
(the last line is needed to ensure the possibility of extracting $x$ from $P$, and thus $\mu$ from $P$), where $y$ ranges over $\mathbb{B}^n$. Choose $x$ satisfying $d_\rho(x) \ll 0$. One can easily verify that the required properties are satisfied. 

The next lemma will be used to prove Theorems 10, 12, and 13.

**Lemma 8.** Let $c > 0$, a random variable $\xi$ satisfy the inequality $0 \leq \xi \leq C$, and $0 \leq I \leq E\xi$. Then $P(\xi \geq I) \geq (E\xi - I)/C$.

**Proof.** Let us denote $p = P\{\xi \geq I\}$. We have
\[
E\xi \leq pc + (1 - p)I = I + p(c - I),
\]
whence $p \geq (E\xi - I)/(c - I) \geq (E\xi - I)/c$.

**Proof of Theorem 10.** Choose an integer $I > 1/\delta$ and rational numbers $\kappa_1, \kappa_2$ such that
\[
(\log e)(2 - 2^{1/2} - \delta) < \kappa_1 < \kappa_2 < (2 - 2^{1/2})(\log e).
\]
Let $n$ be of the form $k^l$, $k \in \mathbb{Z}^\geq$. Let $\mathcal{T}$ be a polynomial-time statistical test which on input $x \in \mathbb{B}^n$, $n = k^l$, outputs 1 if $q_{\pi_n}(x, V_x) \geq \kappa_2 n$ (recall that $V_x$ is defined by $V_x(x) = 1$) and 0 if $q_{\pi_n}(x, V_x) \leq \kappa_1 n$.

First we assume that $x$ is generated in accordance with the uniform $n$-measure $L$. Then
\[
\sum_{x \in \mathbb{B}^n} L(x) q_{\pi_n}(x, V_x) = (\log e) \sum_{x \in \mathbb{B}^n} L(x) \sum_{i=0}^{n-1} \left( \frac{1}{2} \left( (1 - \pi^{1/2})^2 + ((1 - \pi)^{1/2})^2 \right) + \frac{1}{2} \left( (\pi^{1/2})^2 + (1 - (1 - \pi)^{1/2})^2 \right) \right)
\geq (\log e)(2 - 2^{1/2} - (1 - \pi)^{1/2}) \geq (\log e)(2 - 2^{1/2}) n,
\]
where $\pi = \pi_n(x^i)$. Lemma 8 shows that $\mathcal{T}$ outputs 1 with probability separated from 0 (since $q_{\pi_n}$ is bounded from above by $2n$).
Now we assume that $x$ is an output of a CSB generator fed with strings randomly selected from the uniform $k$-measure. For $n$ sufficiently large, $\mathcal{I}$ again outputs 1 with probability separated from 0. Hence, a CSB string $x \in \mathcal{B}^n$ exists such that (16) holds, where $P = V_x$. It is easy to verify that $P$ is an $(n^{1/4} + c, f(n))$-simple $n$-measure and that $d_P(x) \leq 0$. 

10. PROOFS OF THEOREMS 11, 12, 13

In the next lemma the forecasting algorithm $\mathcal{P}(n, \varepsilon, \Gamma, y)$ is constructed.

**Lemma 9.** There are an algorithm $\mathcal{P}$, a constant $c$, and a polynomial $f$ satisfying the following. Let $n, \Gamma \in \mathbb{Z}^+$ and $\varepsilon \in \mathbb{D}$. For $\Gamma \geq c$ the values

$$\pi(y) = \mathcal{P}(n, \varepsilon, \Gamma, y) \in \mathbb{Q}, \quad y \in \mathcal{B}^n,$$

are computed within time $\Gamma$ and comprise an $(n, \varepsilon)$-forecasting system. If $\alpha, \gamma \in \mathbb{Z}^+$ satisfy (6) and an $(n, \varepsilon)$-forecasting system $\mu$ is $(\alpha, \gamma)$-simple, then (7) holds.

**Proof** coincides with the proof of Lemma 3 except that the phrase “yields a rational number in [0, 1]” in the description of stage (ii) of the algorithm must be replaced with “yields a rational number in [\varepsilon, 1 - \varepsilon].” 

**Proof of Theorem 11.** Let us denote for brevity

$$q = q^{\oplus}_n(x, P), \quad d = d^{\oplus}_P(x \mid \Gamma, \mu).$$

Relativizing Lemma 6 (the 1st inequality), we get

$$d_\mu(x \mid \Gamma, P) \leq q + c_\varepsilon^{-\delta} \left( q \log \log q \right)^{1/2} + c \varepsilon^{-\delta} (qd)^{1/2} + cd.$$

As in the proof of Theorem 8 we obtain

$$d_\mu(x \mid \Gamma, \mu, P) \leq q + \varepsilon + 1.01 \log \log q + c \varepsilon^{-\delta} \left( q \log \log q \right)^{1/2}$$

$$+ c \varepsilon^{-\delta} (qd)^{1/2} + c_1 d$$

($c_1, c_2, \ldots$ are integers depending only on $\delta$). Noting that

$$d_\mu(x \mid \Gamma, \mu, \pi) \leq d + c_2,$$

and using (relativized) Lemma 6 (the 2nd inequality), we obtain

$$q_\varepsilon(x, P) \leq q + \varepsilon + 1.01 \log \log q + c \varepsilon^{-\delta} \left( q \log \log q \right)^{1/2} + c \varepsilon^{-\delta} (qd)^{1/2}$$

$$+ c_3 d + c_3 \varepsilon^{-2\delta} \left( q^{1/2} + \alpha^{1/2} + (qd)^{1/4} + d^{1/2} \right)$$

$$\times (\log \log q)^{1/2} + (\log \log \alpha)^{1/2} + (\log \log d)^{1/2}$$

$$+ c_4 \varepsilon^{-2\delta} ((qd)^{1/2} + (\alpha d)^{1/2} + (qd)^{1/4} d^{1/2} + d) + c_5 \varepsilon^{-\delta} d. \quad (31)$$
The underlined terms may be deleted as $(qd)^{1/4} \leq q^{1/2} + d^{1/2}$. From (31) one can easily deduce
\[ q(x, P) \leq (q + \alpha) + c_6 e^{-2\delta ((q + \alpha) \log^\circ \log (q + \alpha))^{1/2}} + c_6 e^{-2\delta ((q + \alpha) d)^{1/2}} + c_6 e^{-2\delta d}. \]

Since
\[ d \leq d_0^\circ(x) + H(\Gamma, \mu | P) + c_7 \]
\[ \leq d_0^\circ(x) + 2K(\Gamma | n) + 2K(\mu | P) + c_8, \]
we get (19).

Proof of Theorem 12. Denote \( m = \lceil n/(1 - \mathcal{H}(\epsilon)) \rceil + \alpha \). First we construct an \((m, \epsilon)\)-measure \( Q \) such that
\[ \sum_{x \in \mathbb{B}^n} Q(x) \rho(Q(1 | x^i), \pi(x^i)) \geq 1 - \mathcal{H}(\epsilon), \quad \forall i < m, \tag{32} \]
and \( Q(1 | x), y \in \mathbb{B}^i \), equals \( \epsilon \) or \( 1 - \epsilon \) and depends only on \( i \).

The measure \( P \) is defined by
\[ P(1 | y) = \begin{cases} Q(1 | y), & \text{if } l(y) < m, \\ 1/2, & \text{otherwise}, \end{cases} \]
where \( y \in \mathbb{B}^n \). Summing over \( i \), we get from (32)
\[ \sum_{x \in \mathbb{B}^n} P(x) F(x) \geq m(1 - \mathcal{H}(\epsilon)), \]
where
\[ F(x) = \sum_{i=0}^{m-1} \rho(P(1 | x^i), \pi(x^i)). \]
Without loss of generality assume that \( \pi \) is an \((n, \epsilon)\)-forecasting system. As
\[ F(x) \leq \sum_{i=0}^{m-1} \left( \epsilon \log \frac{\epsilon}{1 - \epsilon} + (1 - \epsilon) \log \frac{1 - \epsilon}{\epsilon} \right) \]
\[ = m(1 - 2\epsilon) \log((1 - \epsilon)/\epsilon) \leq -m \log \epsilon, \]
we can deduce from Lemma 8
\[ P\{x \in \mathbb{B}^n \mid F(x) \geq (1 + 2^{-\beta} \log \epsilon) m(1 - \mathcal{H}(\epsilon)) \} \]
\[ \geq (-2^{-\beta} \log \epsilon) m(1 - \mathcal{H}(\epsilon))/(-m \log \epsilon) > 2^{-\beta - \delta}. \tag{33} \]
From (14) and (33) one may conclude that there exists $x \in \mathbb{B}^n$ such that

$$F(x) \geq (1 + 2^{-\beta} \log \varepsilon) m(1 - \mathcal{H}(\varepsilon))$$

$$\geq (1 + 2^{-\beta} \log \varepsilon)(\eta + (1 - \mathcal{H}(\varepsilon))\alpha)$$

and $d_p(x) \leq \beta + 3$. It remains to define an $(\alpha + c, f(n))$-simple $n$-forecasting system $\mu$ for which $q_\mu(x, P) \leq \eta + 1$ and $K(\mu | P) \leq c$. Set

$$\mu(y) = \begin{cases} P(1 \mid y), & \text{if } l(y) < \alpha, \\ \frac{1}{2}, & \text{otherwise}, \end{cases}$$

for all $y \in \mathbb{B}^*$. The forecasting system $\mu$ is $(\alpha + c, f(n))$-simple: as its description the sequence $a_0a_1 \cdots a_{\alpha - 1}$ may be taken, where

$$a_i = \begin{cases} 1, & \text{if } \mu(y) = \varepsilon \text{ for } y \in \mathbb{B}', \\ 0, & \text{if } \mu(y) = 1 - \varepsilon \text{ for } y \in \mathbb{B}' \end{cases}$$

(recall that $\varepsilon$ is known to all algorithms). Furthermore,

$$q_\mu(x, P) = \sum_{i=0}^{m-1} \left( \varepsilon \log \frac{\varepsilon}{1/2} + (1 - \varepsilon) \log \frac{1 - \varepsilon}{1/2} \right)$$

$$= (m - \alpha)(1 - \mathcal{H}(\varepsilon)) < \eta + 1. \quad (35)$$

Now a small modification of $P$ will insure that, on the one hand, the inequalities between the extreme terms of (33) and (35) continue to hold, and, on the other hand, one is able to effectively extract $\alpha$ from $P$. Since $\mu$ can be effectively found given $P$ and $\alpha$ (though (34) may be violated after the modification of $P$), it follows that $K(\mu | P) \leq c$. 

**Proof of Theorem 13.** Choose an integer $l > 1/\delta$ and rational numbers $\kappa_1, \kappa_2$ such that

$$1 - \mathcal{H}(\varepsilon) + 2^{-\beta} \log \varepsilon < \kappa_1 < \kappa_2 < 1 - \mathcal{H}(\varepsilon) + 2^{-\beta - 1/2} \log \varepsilon.$$

Let $n$ be of the form $k', k \in \mathbb{Z}^\oplus$. We define $\mathcal{I}$ as an arbitrary polynomial-time statistical test $\mathcal{I}$ satisfying what follows. Receiving as input a string $z \in \mathbb{B}^n$, $n = k'$, $\mathcal{I}$ randomly selects a string $x$ in accordance with the measure $Q_z$ defined by the equality

$$Q_z(1 \mid y) = \begin{cases} 1 - \varepsilon, & \text{if } z_i = 1, \\ \varepsilon, & \text{if } z_i = 0, \end{cases}$$

where $i < n$, $y \in \mathbb{B}'$. Then $\mathcal{I}$ outputs 1, if $q_{\pi_n}(x, Q_z) \geq \kappa_2n$, and 0, if $q_{\pi_n}(x, Q_z) \leq \kappa_1n$. 

Let $z$ be randomly selected from the uniform $n$-measure $L$. We have
\[
\sum_{z \in \mathbb{B}^n} L(z) \sum_{x \in \mathbb{B}^n} Q_z(x) q_{\pi_{n,\varepsilon}}(x, Q_z)
\]
\[
= \sum_{z \in \mathbb{B}^n} L(z) \sum_{x \in \mathbb{B}^n} Q_z(x) \sum_{i=0}^{n-1} \left( \frac{1}{2} \left( (1 - \varepsilon) \log \frac{1 - \varepsilon}{\pi} + \varepsilon \log \frac{\varepsilon}{1 - \pi} \right) \right)
\]
\[
\geq -nH(\varepsilon) + n = (1 - H(\varepsilon))n,
\]
where $\pi = \pi_{n,\varepsilon}(x^i)$. As in the proof of Theorem 12 one can deduce from Lemma 8 that $\mathcal{I}$ outputs 1 with probability no less than $2^{-\beta - 1/2}$ (we can again assume that $\pi_{n,\varepsilon}$ is an $(n, \varepsilon)$-forecasting system).

Now let $z$ be an output of a CSB generator fed with strings randomly selected from the uniform $k$-measure. From some $n$ on, $\mathcal{I}$ outputs 1 with probability greater than $2^{-\beta - 1/2}$; therefore, for some CSB string $z$ the set of $x$ satisfying $q_{\pi_{n,\varepsilon}}(x, Q_z) > \kappa_1 n$ has $Q_z$-measure greater than $2^{-\beta - 1}$. It is clear from (14) that a string $x$ exists such that (20) and $d_P(x) \leq \beta + 1$ hold, where $P = Q_z$. Clearly, $P$ is an $(n^{1/2} + c, f(n))$-simple $(n, \varepsilon)$-measure.

ACKNOWLEDGMENTS

I thank the referee for a detailed and critical review with suggestions for strengthening some results and improvements in the presentation. I am grateful to Professor G. Mauri for providing information about (Littlestone, Warmuth, 1989).

RECEIVED January 29, 1989; FINAL MANUSCRIPT RECEIVED June 25, 1990

REFERENCES


