On Calibration Error of Randomized Forecasting Algorithms

Vladimir V. V'yugin

Institute for Information Transmission Problems, Russian Academy of Sciences,
Bol'shoi Karetnyi per. 19, Moscow GSP-4, 127994, Russia
vyugin@iitp.ru

Abstract. Recently, it was shown that calibration with an error less than $\delta > 0$ is almost surely guaranteed with a randomized forecasting algorithm, where forecasts are chosen using randomized rounding up to δ of deterministic forecasts. We show that this error can not be improved for a large majority of sequences generated by a probabilistic algorithm: we prove that combining outcomes of coin-tossing and a transducer algorithm, it is possible to effectively generate with probability close to one a sequence "resistant" to any randomized rounding forecasting with an error much smaller than δ.

1 Introduction

A minimal requirement for testing of any prediction algorithm is that it should be calibrated (see Dawid [1]). An informal explanation of calibration would go something like this. Let a binary sequence $\omega_1, \omega_2, \ldots, \omega_{n-1}$ of outcomes is observed by a forecaster whose task is to give a probability p_n of a future event $\omega_n = 1$. A typical example is that p_n is interpreted as a probability that it will rain. Forecaster is said to be well-calibrated if it rains as often as he leads us to expect. It should rain about 80% of the days for which $p_n = 0.8$, and so on. So, for simplicity we consider binary sequences, i.e. $\omega_n \in \{0, 1\}$ for all n. We give a rigorous definition of calibration later.

We suppose that the forecasts p_n are computed by some algorithm. If the weather acts adversarially, then Oakes [6] and Dawid [2] show that a deterministic forecasting algorithm will not be always be calibrated. V'yugin [9] proved that this result holds for a large majority of sequences generated by a probabilistic algorithm.

Foster and Vohra [4] show that calibration is almost surely guaranteed with a randomizing forecasting rule, i.e., where the forecasts are chosen using private randomization and the forecasts are hidden from the weather until weather makes its decision to rain or not. Kakade and Foster [5] obtained an analogous positive result for deterministic forecasting systems and for the case where the class of "selection rules" is restricted to "continuous selection rules". This approach was further developed in Vovk et al. [8].

In Section 2 we give the definition of calibration and randomized rounding. Main result of this paper - Theorem 1, is presented in Section 3, the proof of the

M. Hutter, R.A. Servedio, and E. Takimoto (Eds.): ALT 2007, LNAI 4754, pp. 380–394, 2007.

main result is given in Section 4. This theorem shows that combining outcomes of coin-tossing and a transducer algorithm, it is possible to effectively generate with probability close to one a sequence "resistant" to randomized rounding forecasting with error much smaller than the precision of rounding. Theorems 2 and 3 show that the calibration error may be much bigger if we check calibration using "deterministic selection rules".

2 Background

Let Ω be the set of all infinite binary sequences, Ξ be the set of all finite binary sequences, and λ be the empty sequence. For any finite or an infinite sequence $\omega = \omega_1 \ldots \omega_n \ldots$ we write $\omega^n = \omega_1 \ldots \omega_n$ (we put $\omega_0 = \omega^0 = \lambda$). Also, $l(\omega^n) = n$ denotes the length of the sequence ω^n. If x is a finite sequence and ω is a finite or infinite sequence then $x\omega$ denotes the concatenation of these sequences, $x \sqsubseteq \omega$ means that $x = \omega^n$ for some n.

A deterministic *forecasting system* f is a real-valued function $f : \Xi \to [0, 1]$. We consider computable forecasting systems; there is an algorithm, which given a finite sequence $\omega_1 \ldots \omega_{n-1} \in \Xi$ and an arbitrary positive rational number κ, when halts, outputs a rational approximation of $f(\omega_1 \ldots \omega_{n-1})$ up to κ. A forecasting system f is called *total* if it is defined on each finite sequence $\omega_1 \ldots \omega_{n-1}$. Any total forecasting system defines the corresponding overall probability distribution P on the set of all sequences such that its conditional probabilities satisfy

$$p_n = P(\omega_n = 1 | \omega_1, \omega_2, \ldots, \omega_{n-1}),$$

where $p_n = f(\omega_1 \ldots \omega_{n-1})$. In the following we consider only total forecasting systems.

The evaluation of probability forecasts is based on a method called *calibration* (see Dawid [1], [2]). Let f be some forecasting system and $I(p)$ be a characteristic function of some subinterval $I \subseteq [0, 1]$, i.e., $I(p) = 1$ if $p \in I$, and $I(p) = 0$, otherwise. Let $\omega = \omega_1\omega_2 \ldots$ be an infinite binary sequence.

A forecasting system f is well-calibrated for an infinite sequence $\omega_1\omega_2 \ldots$ if for the characteristic function $I(p)$ of any subinterval of $[0, 1]$ *the calibration error* tends to zero, i.e.,

$$\frac{\sum_{i=1}^{n} I(p_i)(\omega_i - p_i)}{\sum_{i=1}^{n} I(p_i)} \longrightarrow 0 \tag{1}$$

as the denominator of the relation (1) tends to infinity; we denote $p_i = f(\omega^{i-1})$. Here, $I(p_i)$ determines some "selection rule" which define moments of time where we compute the deviation between forecasts p_i and outcomes ω_i.

Oakes [6] proposed arguments (see Dawid [3] for different proof) that no deterministic forecasting system can be well-calibrated for all possible sequences: any total forecasting system f is not calibrated for the sequence $\omega = \omega_1\omega_2 \ldots$, where

$$\omega_i = \begin{cases} 1 \text{ if } p_i < 0.5 \\ 0 \text{ otherwise} \end{cases}$$

and $p_i = f(\omega_1 \dots \omega_{i-1})$, $i = 1, 2, \dots$.

A randomized forecasting system $f(\omega^{n-1})$ is a random variable with range in $[0, 1]$ defined on some probability space supplied with a probability distribution Pr_n, where $\omega^{n-1} \in \Xi$ is a parameter of this variable. For any n, the predictor chooses the forecast p_n of the event $\omega_n = 1$ randomly using probability distribution Pr_n of the variable $f(\omega^{n-1})$. In this case, for any given ω we can consider the probability Pr of the event (1), where Pr is the overall probability distribution generated by probability distributions Pr_n, $n = 1, 2, \dots$.

In the following we suppose that for any ω^{n-1} the range of the random variable $f(\omega^{n-1})$ is finite, say, $\{p_{n,1}, \dots, p_{n,m_n}\}$. The number

$$\delta_n = \inf\{|p_{n,i} - p_{n,j}| : i \neq j\}$$

is called *the level of discreteness of f on ω^{n-1}*. We also consider $\delta = \inf_n \delta_n$ - the level of discretness of f on ω.

A typical example is the uniform rounding: for any n the rational points $p_{n,i}$ divide the unit interval into equal parts of size $0 < \delta < 1$; then the level of discreteness is constant and equals δ.

Kakade and Foster [5] presented "an almost deterministic" *randomized rounding* total forecasting algorithm f: an observer can only randomly round with the precision of rounding (level of discreteness) δ the deterministic forecast in order to calibrate. Then for any infinite sequence $\omega = \omega_1\omega_2 \dots$ the overall probability Pr of the event

$$\left| \frac{1}{n} \sum_{i=1}^{n} I(p_i)(\omega_i - p_i) \right| \leq \delta$$

tends to one as $n \to \infty$, where p_i is the random variable $f(\omega^{n-1})$, $I(p)$ is the characteristic function of any subinterval of $[0, 1]$.[1] This algorithm randomly rounds a forecast computed by some deterministic algorithm (constructed in [5]): for example, the forecast 0.8512 can be rounded up to second digit to 0.86 with probability 0.12, and to 0.85 with probability 0.88, at the next moment of time, the forecast 0.2588 can be rounded up to second digit to 0.26 with probability 0.88, and to 0.25 with probability 0.12. Here we have in mind some algorithm defining the direction of rounding.

[1] In fact, more accurate calculations show that this inequality can be replaced on

$$\left| \frac{1}{\alpha(n)\sqrt{n}} \sum_{i=1}^{n} I(p_i)(\omega_i - p_i) \right| \leq \delta,$$

where $\alpha(n)$ is any unbounded nondecreasing function; then (1) holds for Kakade and Foster's algorithm if $\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(p_i) = \infty$. We do not go into details, since in this paper we prove results in the opposite direction (see [5], [8]).

3 Main Results

We need some computability concepts. Let $\mathcal{R}$ be the set of all real numbers extended by adding the infinities $-\infty$ and $+\infty$, A is some set of finite objects; the elements of A can be effectively enumerated by positive integer numbers (see Rogers [7]). In particular, we will identify a computer program and its number. We fix some effective one-to-one enumeration of all pairs (triples, and so on) of nonnegative integer numbers. We identify any pair (t, s) and its number $\langle t, s\rangle$.

A function $\phi\colon A \to \mathcal{R}$ is called (lower) semicomputable if $\{(r, x) : r < \phi(x)\}$ (r is a rational number) is a recursively enumerable set. This means that there is an algorithm which when fed with a rational number r and a finite object x eventually stops if $r < \phi(x)$ and never stops, otherwise. In other words, the semicomputability of f means that if $\phi(x) > r$ this fact will sooner or later be learned, whereas if $f(x) \le r$ we may be for ever uncertain. A function ϕ is upper semicomputable if $-\phi$ is lower semicomputable.

Standard argument based on the recursion theory shows that there exist the lower and upper semicomputable real functions $\phi^-(j, x)$ and $\phi^+(k, x)$ universal for all lower semicomputable and upper semicomputable functions from $x \in \Xi$.[2] As follows from the definition, for every computable real function $\phi(x)$ there exist a pair $\langle j, k\rangle$ such that

$$\phi(x) = \phi^-(j, x) = \phi^+(k, x)$$

for all x. Let $\phi_s^-(j, x)$ be equal to the maximal rational number r such that the triple (r, j, x) is enumerated in s steps in the process of enumerating of the set

$$\{(r, j, x) : r < \phi(j, x),\ r \text{ is rational}\}$$

and equals $-\infty$, otherwise. Any such function $\phi_s^-(j, x)$ takes only finite number of rational values distinct from $-\infty$. By definition, $\phi_s^-(j, x) \le \phi_{s+1}^-(j, x)$ for all j, s, x, and

$$\phi^-(j, x) = \lim_{s\to\infty} \phi_s^-(j, x).$$

An analogous non-increasing sequence of functions $\phi_s^+(k, x)$ exists for any upper semicomputable function.

Let $i = \langle t, k\rangle$. We say that the function $\phi_i(x)$ is *defined on* x if given any degree of precision - positive rational number $\kappa > 0$, it holds

$$|\phi_s^+(t, x) - \phi_s^-(k, x)| \le \kappa$$

for some s; $\phi_i(x)$ undefined, otherwise. If any such s exists then for minimal such s, $\phi_{i,\kappa}(x) = \phi_s^-(k, x)$ is called the rational approximation (from below) of $\phi_i(x)$ up to κ; $\phi_{i,\kappa}(x)$ undefined, otherwise.

Any measure P on Ω can be defined as follows. Let us consider intervals

$$\Gamma_z = \{\omega \in \Omega : z \sqsubseteq \omega\},$$

[2] This means that each lower semicomputable function $\phi(x)$ can be represented as $\phi(x) = \phi^-(j, x)$ for some j. The same holds for upper semicomputability.

where $z \in \Xi$. We denote $P(z) = P(\Gamma_z)$ for $z \in \Xi$ and extend this function on all Borel subsets of Ω in a standard way.

A measure P is computable if there exists an algorithm which given $z \in \Xi$ and a degree of precision κ computes the number $P(z)$ up to κ.

We use also a concept of *computable operation* on $\Xi \bigcup \Omega$ [10,11]. Let $\hat{F}$ be a recursively enumerable set of ordered pairs of finite sequences satisfying the following properties:

- (i) $(x, \lambda) \in \hat{F}$ for each x;
- (ii) if $(x, y) \in \hat{F}$, $(x', y') \in \hat{F}$ and $x \sqsubseteq x'$ then $y \sqsubseteq y'$ or $y' \sqsubseteq y$ for all finite binary sequences x, x', y, y'.

A computable operation F is defined as follows

$$F(\omega) = \sup\{y \mid x \sqsubseteq \omega \text{ and } (x, y) \in \hat{F} \text{ for some } x\},$$

where $\omega \in \Omega \bigcup \Xi$ and sup is in the sense of the partial order $\sqsubseteq$ on Ξ.

Informally, the computable operation F is defined by some algorithm; this algorithm when fed with an infinite or a finite sequence ω takes it sequentially bit by bit, processes it, and produces an output sequence also sequentially bit by bit.

A probabilistic algorithm is a pair (P, F), where P is a computable measure on the set of all binary sequences and F is a computable operation. For any probabilistic algorithm (P, F) and a set $A \subseteq \Omega$, we consider the probability

$$P\{\omega : F(\omega) \in A\}$$

of generating by means of F a sequence from A given a sequence ω distributed according to the computable probability distribution P. In the following $P = L$, where $L(x) = L(\Gamma_x) = 2^{-l(x)}$ is the uniform measure on Ω.

A natural definition of computable randomized forecasting system f would be the following: a random variable f is computable if its probability distribution function

$$\phi(\alpha; \omega^{n-1}) = Pr_n\{f(\omega^{n-1}) < \alpha\}$$

is "a computable real function" from arguments $\alpha \in [0, 1]$ and $\omega^{n-1} \in \Xi$. The precise definition requires some technicalities. In fact, in the construction below, we compute ϕ only at one point $\alpha = 0.5$; so, we will use the following definition. A randomized forecasting system f is *weakly computable* if its *weak probability distribution function*

$$\varphi_n(\omega^{n-1}) = Pr_n\{f(\omega^{n-1}) < 0.5\}$$

is a computable function from ω^{n-1}.

Let $I_0 = I_0(p)$ be the characteristic function of the interval $(0, \frac{1}{2})$ and $I_1 = I_1(p)$ be the characteristic function of the interval $[\frac{1}{2}, 1)$. The following theorem is the main result of this paper.

Theorem 1. *For any $\epsilon > 0$ a probabilistic algorithm (L, F) can be constructed, which with probability $\geq 1 - \epsilon$ outputs an infinite binary sequence $\omega = \omega_1\omega_2\ldots$ such that for every weakly computable randomized forecasting system f with level of discretness δ on ω, for some $\nu = 0$ or $\nu = 1$, the overall probability of the event*

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I_\nu(p_i)(\omega_i - p_i) \right| \geq 0.25\delta \tag{2}$$

equals one, where the overall probability is associated with f and $p_i = f(\omega^{i-1})$, $i = 1, 2, \ldots$, is a random variable.

The following deterministic analogue of Theorem 1 was obtained in V'yugin [9].

Theorem 2. *For any $\epsilon > 0$ a probabilistic algorithm (L, F) can be constructed, which with probability $\geq 1 - \epsilon$ outputs an infinite binary sequence $\omega = \omega_1\omega_2\ldots$ such that for every deterministic forecasting algorithm f, for some $\nu = 0$ or $\nu = 1$,*

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I_\nu(p_i)(\omega_i - p_i) \right| \geq 0.5,$$

where $p_i = f(\omega^{i-1})$, $i = 1, 2, \ldots$.

Theorem 1 uses randomized "selection rules" - $I_\nu(p_i)$, $\nu = 0, 1$. In case of some natural deterministic "selection rule", we obtain the following theorem.

Let $E(f(\omega^{i-1})$ be the mean value of the forecasts produced by a randomized forecasting system f given an input sequence ω^{i-1}.

In the following theorem we use more strong definition of computability of randomized forecasting systems - we consider randomized forecasting systems with computable mathematical expectations: for any such system f its mathematical expectation $E(f(\omega^{i-1})$ is a computable real function from ω^{n-1}.

Theorem 3. *For any $\epsilon > 0$ a probabilistic algorithm (L, F) can be constructed, which with probability $\geq 1 - \epsilon$ outputs an infinite binary sequence $\omega = \omega_1\omega_2\ldots$ such that for every randomized forecasting system f with computable mathematical expectation, for some $\nu = 0$ or $\nu = 1$, the overall probability of the event*

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} I_\nu(E(p_i))(\omega_i - p_i) \right| \geq 0.5 \tag{3}$$

equals one, where $p_i = f(\omega^{i-1})$, $i = 1, 2, \ldots$.

4 Proofs of Theorems 1-3

For any probabilistic algorithm (P, F), we consider the function

$$Q(x) = P\{\omega : x \sqsubseteq F(\omega)\}. \tag{4}$$

It is easy to verify that this function is lower semicomputable and satisfies:

$$Q(\lambda) \leq 1;$$
$$Q(x0) + Q(x1) \leq Q(x)$$

for all x. Any function satisfying these properties is called semicomputable semimeasure. For any semicomputable semimeasure Q a probabilistic algorithm (L, F) exists such that (4) holds, where $P = L$ (for the proof see [10,11]).

Though the semimeasure Q is not a measure, we consider the corresponding measure on the set Ω

$$\bar{Q}(\Gamma_x) = \inf_n \sum_{l(y)=n, x \sqsubseteq y} Q(y).$$

This function can be extended on all Borel subsets A of Ω (see [11]).

We will construct a semicomputable semimeasure Q as a some sort of network flow. We define an infinite network on the base of the infinite binary tree. Any $x \in \Xi$ defines two edges $(x, x0)$ and $(x, x1)$ of length one. In the construction below we will mount to the network extra edges (x, y) of length > 1, where $x, y \in \Xi$, $x \sqsubseteq y$ and $y \neq x0, x1$. By the length of the edge (x, y) we mean the number $l(y) - l(x)$. For any edge $\sigma = (x, y)$ we denote by $\sigma_1 = x$ its starting vertex and by $\sigma_2 = y$ its terminal vertex. A computable function $q(\sigma)$ defined on all edges of length one and on all extra edges and taking rational values is called *a network* if for all $x \in \Xi$

$$\sum_{\sigma:\sigma_1=x} q(\sigma) \leq 1.$$

Let G be the set of all extra edges of the network q (it is a part of the domain of q). By *q-flow* we mean the minimal semimeasure P such that $P \geq R$, where the function R is defined by the following recursive equations

$$R(\lambda) = 1;$$
$$R(y) = \sum_{\sigma:\sigma_2=y} q(\sigma) R(\sigma_1) \tag{5}$$

for $y \neq \lambda$. We can define the semimeasure P using R as follows. A set D is prefix-free if $x \not\sqsubseteq y$ for all $x, y \in D$, $x \neq y$. Let π_x be the set of all prefix-free sets D such that $x \sqsubseteq y$ for all $y \in D$. Then it holds

$$P(x) = \sup_{D \in \pi_x} \sum_{x:x \in D} R(x).$$

A network q is called *elementary* if the set of extra edges is finite and $q(\sigma) = 1/2$ for almost all edges of unit length. For any network q, we define the *network flow delay* function (q-delay function)

$$d(x) = 1 - q(x, x0) - q(x, x1).$$

The construction below works with all computable real functions $\phi_t(x)$, $x \in \Xi$, $t = 1, 2, \ldots$; any $i = \langle t, s\rangle$ is considered as a program for computing the rational approximation $\phi_{t,\kappa_s}(\omega^{n-1})$ of ϕ_t from below up to $\kappa_s = 1/s$.[3] In the proof (see Lemma 6) we use a special class of these functions, namely, functions of the type

$$\phi(\omega^{n-1}) = Pr_n\{f(\omega^{n-1}) \geq 0.5\} = 1 - \varphi_n(\omega^{n-1}), \tag{6}$$

where $\varphi_n(\omega^{n-1})$ is a weak probability distribution function for some weakly computable randomized forecasting system f. We have $\phi = \phi_t$ for some t, and by the construction below we visit any function ϕ_t on infinitely many steps n. To do this, we define some function $p(n)$ such that for any positive integer number i we have $p(n) = i$ for infinitely many n. For example, we can define $p(\langle i, s\rangle) = i$ for all i and s.

For any program $i = \langle t, s\rangle$, any finite binary sequences x and y, any elementary network q, and for any integer number n, let $B(i, x, y, q, n)$ be *true* if the following conditions hold

- (i) $l(y) = n$, $x \sqsubseteq y$,
- (ii) $d(y^k) < 1$ for all k, $1 \leq k \leq n$, where d is the q-delay function and $y^k = y_1 \ldots y_k$;
- (iii) for all k such that $l(x) \leq k < il(x)$ the values $\phi_{t,\kappa_s}(y^k)$ are defined in $\leq n$ steps and

$$y_{k+1} = \begin{cases} 0 \text{ if } \phi_{t,\kappa_s}(y^k) \geq 0.5 \\ 1 \text{ otherwise.} \end{cases}$$

Let $B(i, x, y, q, n)$ be *false*, otherwise. Define

$$\beta(x, q, n) = \min\{y : p(l(y)) = p(l(x)), B(p(l(x)), x, y, q, n)\}$$

Here min is considered for lexicographical ordering of strings; we suppose that $\min \emptyset$ is undefined.

Lemma 1. *For any total function ϕ_t, $\beta(x, q, n)$ is defined for all $x \in \Xi$ and for all sufficiently large n.*

Proof. The needed sequence y can be easily defined for all sufficiently large n sequentially bit-by-bit, since $\phi_{t,\kappa_s}(z)$ is defined for all z. □

The goal of the construction below is the following. Any extra edge σ will be assigned to some task number i such that $p(l(\sigma_1)) = p(l(\sigma_2)) = i$. The goal of the task i is to define a finite set of extra edges σ such that for any infinite binary sequence ω one of the following conditions hold: either ω contains some extra edge as a subword, or the network flow delay function d equals 1 on some initial fragment of ω. For any extra edge σ mounted to the network q, $B(i, \sigma_1, \sigma_2, q^{n-1}, n)$ is true; it is false, otherwise. Lemma 5 shows that $\bar{Q}(E_Q) > 1 - \frac{1}{2}\epsilon$, where Q is

[3] Recall that $t = \langle j, k\rangle$ for some j, k; we use the lower and upper semicomputable real functions $\phi^-(j, x)$ and $\phi^+(k, x)$ universal for all lower semicomputable and upper semicomputable functions from $x \in \Xi$ to compute values $\phi_t(x)$.

the q-flow and E_Q is defined by (7) below. Lemma 6 shows that for each $\omega \in E_Q$ the event (2) holds with the overall probability one.

Construction. Let $\rho(n) = (n+n_0)^2$ for some sufficiently large n_0 (the value n_0 will be specified below in the proof of Lemma 5).

Using the mathematical induction by n, we define a sequence q^n of elementary networks. Put $q^0(\sigma) = 1/2$ for all edges σ of length one.

Let $n > 0$ and a network q^{n-1} is defined. Let d^{n-1} be the q^{n-1}-delay function and let G^{n-1} be the set of all extra edges. We suppose also that $l(\sigma_2) < n$ for all $\sigma \in G^{n-1}$.

Let us define a network q^n. At first, we define a network flow delay function d^n and a set G^n. The construction can be split up into two cases.

Let $w(i, q^{n-1})$ be equal to the minimal m such that $p(m) = i$ and $m > (i+1)l(\sigma_2)$ for each extra edge $\sigma \in G^{n-1}$ such that $p(l(\sigma_1))) < i$.

The inequality $w(i, q^m) \neq w(i, q^{m-1})$ can be induced by some task $j < i$ that mounts an extra edge $\sigma = (x, y)$ such that $l(y) > w(i, q^{m-1})$ and $p(l(x)) = p(l(y)) = j$. Lemma 2 (below) will show that this can happen only at finitely many steps of the construction.

Case 1. $w(p(n), q^{n-1}) = n$ (the goal of this part is to start a new task $i = p(n)$ or to restart the existing task $i = p(n)$ if it was destroyed by some task $j < i$ at some preceding step).

Put $d^n(y) = 1/\rho(n)$ for $l(y) = n$ and define $d^n(y) = d^{n-1}(y)$ for all other y. Put also $G^n = G^{n-1}$.

Case 2. $w(p(n), q^{n-1}) < n$ (the goal of this part is to process the task $i = p(n)$).

Let C_n be the set of all x such that $w(i, q^{n-1}) \leq l(x) < n$, $0 < d^{n-1}(x) < 1$, the function $\beta(x, q^{n-1}, n)$ is defined[4] and there is no extra edge $\sigma \in G^{n-1}$ such that $\sigma_1 = x$.

In this case for each $x \in C_n$ define $d^n(\beta(x, q^{n-1}, n)) = 0$, and for all other y of length n such that $x \sqsubset y$ define

$$d^n(y) = \frac{d^{n-1}(x)}{1 - d^{n-1}(x)}.$$

Define $d^n(y) = d^{n-1}(y)$ for all other y. We add an extra edge to G^{n-1}, namely, define

$$G^n = G^{n-1} \cup \{(x, \beta(x, q^{n-1}, n)) : x \in C_n\}.$$

We say that the task $i = p(n)$ *mounts* the extra edge $(x, \beta(x, q^{n-1}, n))$ to the network and that all existing tasks $j > i$ are destroyed by the task i.

After Case 1 and Case 2, define for any edge σ of unit length

$$q^n(\sigma) = \frac{1}{2}(1 - d^n(\sigma_1))$$

and $q^n(\sigma) = d^n(\sigma_1)$ for each extra edge $\sigma \in G^n$.

[4] In particular, $p(l(x)) = i$ and $l(\beta(x, q^{n-1}, n)) = n$.

Case 3. Cases 1 and 2 do not hold.

Define $d^n = d^{n-1}$, $q^n = q^{n-1}$, $G^n = G^{n-1}$.

As the result of the construction we define the network $q = \lim\limits_{n\to\infty} q^n$, the network flow delay function $d = \lim\limits_{n\to\infty} d^n$ and the set of extra edges $G = \cup_n G^n$.

The functions q and d are computable and the set G is recursive by their definitions. Let Q denotes the q-flow.

The following lemma shows that any task can mount new extra edges only at finite number of steps. Let $G(i)$ be the set of all extra edges mounted by the task i, $w(i,q) = \lim_{n\to\infty} w(i,q^n)$.

Lemma 2. *The set $G(i)$ is finite, $w(i,q)$ exists and $w(i,q) < \infty$ for all i.*

Proof. Note that if $G(j)$ is finite for all $j < i$, then $w(i,q) < \infty$. Hence, we must prove that the set $G(i)$ is finite for any i. Suppose that the opposite assertion holds. Let i be the minimal such that $G(i)$ is infinite. By choice of i the sets $G(j)$ for all $j < i$ are finite. Then $w(i,q) < \infty$.

For any x such that $l(x) \geq w(i,q)$, consider the maximal m such that for some initial fragment $x^m \sqsubseteq x$ there exists an extra edge $\sigma = (x^m, y) \in G(i)$. If no such extra edge exists define $m = w(i,q)$. By definition, if $d(x^m) \neq 0$ then $1/d(x^m)$ is an integer number. Define

$$u(x) = \begin{cases} 1/d(x^m) \text{ if } d(x^m) \neq 0, l(x) \geq w(i,q) \\ \rho(w(i,q)) \text{ if } l(x) < w(i,q) \\ 0 \text{ otherwise} \end{cases}$$

By construction the integer valued function $u(x)$ has the property: $u(x) \geq u(y)$ if $x \sqsubseteq y$. Besides, if $u(x) > u(y)$ then $u(x) > u(z)$ for all z such that $x \sqsubseteq z$ and $l(z) = l(y)$. Then the function

$$\hat{u}(\omega) = \min\{n : u(\omega^i) = u(\omega^n) \text{ for all } i \geq n\}$$

is defined for all $\omega \in \Omega$. It is easy to see that this function is continuous. Since Ω is compact space in the topology generated by intervals Γ_x, this function is bounded by some number m. Then $u(x) = u(x^m)$ for all $l(x) \geq m$. By the construction, if any extra edge of ith type was mounted to $G(i)$ at some step then $u(y) < u(x)$ holds for some new pair (x,y) such that $x \sqsubseteq y$. This is contradiction with the existence of the number m. □

An infinite sequence $\alpha \in \Omega$ is called an *i-extension* of a finite sequence x if $x \sqsubseteq \alpha$ and $B(i, x, \alpha^n, n)$ is true for almost all n.

A sequence $\alpha \in \Omega$ is called *i-closed* if $d(\alpha^n) = 1$ for some n such that $p(n) = i$, where d is the q-delay function. Note that if $\sigma \in G(i)$ is some extra edge then $B(i, \sigma_1, \sigma_2, n)$ is true, where $n = l(\sigma_2)$.

Lemma 3. *Let for any initial fragment ω^n of an infinite sequence ω some i-extension exists. Then either the sequence ω will be i-closed in the process of the construction or ω contains an extra edge of ith type (i.e. $\sigma_2 \sqsubseteq \omega$ for some $\sigma \in G(i)$).*

Proof. Let a sequence ω is not i-closed. By Lemma 2 the maximal m exists such that $p(m) = i$ and $d(\omega^m) > 0$. Since the sequence ω^m has an i-extension and $d(\omega^k) < 1$ for all k, by Case 2 of the construction a new extra edge (ω^m, y) of ith type must be mounted to the binary tree. By the construction $d(y) = 0$ and $d(z) \neq 0$ for all z such that $\omega^m \sqsubseteq z$, $l(z) = l(y)$, and $z \neq y$. By the choice of m we have $y \sqsubseteq \omega$. □

Lemma 4. *It holds $Q(y) = 0$ if and only if $q(\sigma) = 0$ for some edge σ of unit length located on y (this edge satisfies $\sigma_2 \sqsubseteq y$).*

Proof. The necessary condition is obvious. To prove that this condition is sufficient, let us suppose that $q(y^n, y^{n+1}) = 0$ for some $n < l(y)$ but $Q(y) \neq 0$. Then by definition $d(y^n) = 1$. Since $Q(y) \neq 0$ an extra edge $(x, z) \in G$ exists such that $x \sqsubseteq y^n$ and $y^{n+1} \sqsubseteq z$. But, by the construction, this extra edge can not be mounted to the network $q^{l(z)-1}$ since $d(z^n) = 1$. This contradiction proves the lemma. □

For any semimeasure P define

$$E_P = \{\omega \in \Omega : \forall n(P(\omega^n) \neq 0)\}$$

the support set of P. It is easy to see that E_P is a closed subset of Ω and $\bar{P}(E_P) = \bar{P}(\Omega)$. By Lemma 4, the relation $Q(y) = 0$ is recursive and

$$E_Q = \Omega \setminus \cup_{d(x)=1} \Gamma_x. \tag{7}$$

Lemma 5. *It holds $\bar{Q}(E_Q) > 1 - \frac{1}{2}\epsilon$.*

Proof. We bound $\bar{Q}(\Omega)$ from below. Let R be defined by (5). By definition of the network flow delay function, we have

$$\sum_{u:l(u)=n+1} R(u) = \sum_{u:l(u)=n} (1 - d(u))R(u) + \sum_{\sigma:\sigma\in G, l(\sigma_2)=n+1} q(\sigma)R(\sigma_1). \tag{8}$$

Define an auxiliary sequence

$$S_n = \sum_{u:l(u)=n} R(u) - \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} q(\sigma)R(\sigma_1).$$

At first, we consider the case $w(p(n), q^{n-1}) < n$. If there is no edge $\sigma \in G$ such that $l(\sigma_2) = n$ then $S_{n+1} \geq S_n$. Suppose that some such edge exists. Define

$$P(u, \sigma) \Longleftrightarrow l(u) = l(\sigma_2) \& \sigma_1 \sqsubseteq u \& u \neq \sigma_2 \& \sigma \in G.$$

By definition of the network flow delay function, we have

$$\sum_{u:l(u)=n} d(u)R(u) = \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} d(\sigma_2) \sum_{u:P(u,\sigma)} R(u) =$$

$$= \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} \frac{d(\sigma_1)}{1 - d(\sigma_1)} \sum_{u:P(u,\sigma)} R(u) \leq \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} d(\sigma_1)R(\sigma_1) =$$

$$= \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} q(\sigma)R(\sigma_1). \tag{9}$$

Here we used the inequality

$$\sum_{u:P(u,\sigma)} R(u) \le R(\sigma_1) - d(\sigma_1)R(\sigma_1)$$

for all $\sigma \in G$ such that $l(\sigma_2) = n$. Combining this bound with (8) we obtain $S_{n+1} \ge S_n$.

Let us consider the case $w(p(n), q^{n-1}) = n$. Then

$$\sum_{u:l(u)=n} d(u)R(u) \le \rho(n) = \frac{1}{(n+n_0)^2}.$$

Combining (8) and (9) we obtain

$$S_{n+1} \ge S_n - \frac{1}{(n+n_0)^2}$$

for all n. Since $S_0 = 1$, this implies

$$S_n \ge 1 - \sum_{i=1}^{\infty} \frac{1}{(i+n_0)^2} \ge 1 - \frac{1}{2}\epsilon$$

for some sufficiently large constant n_0. Since $Q \ge R$, it holds

$$\bar{Q}(\Omega) = \inf_n \sum_{l(u)=n} Q(u) \ge \inf_n S_n \ge 1 - \frac{1}{2}\epsilon.$$

Lemma is proved. □

Lemma 6. *For each weakly computable randomized forecasting system f and and for each sequence $\omega \in E_Q$, the event (2) holds with the overall probability one.*

Proof. Let ω be an infinite sequence and let f be a weakly computable randomized forecasting system, i.e., the corresponding $\phi_t(\omega^{n-1})$ (defined by (6)) is defined for all n. Let $i = \langle t, s\rangle$ be a program for computing the rational approximation ϕ_{t,κ_s} from below up to $\kappa_s = 1/s$. Since in the construction we visit ϕ_t on infinitely many steps n such that $p(n) = i = \langle t, s\rangle$, where $s = 1, 2, \ldots$, in the proof we will consider only sufficiently large i.

By definition $d(\omega^n) < 1$ for all n if $\omega \in E_Q$. Since ω is an i-extension of ω^n for each n, by Lemma 3 there exists an extra edge $\sigma \in G(i)$ such that $\sigma_2 \sqsubseteq \omega$. In the following, let $k = l(\sigma_1)$ and $n = ik$.

Denote $p_j^- = \max\{p_{j,s} : p_{j,s} < 0.5\}$ and $p_j^+ = \min\{p_{j,s} : p_{j,s} \ge 0.5\}$, where $\{p_{j,1}, \ldots, p_{j,m_j}\}$ is the range of the random variable $f(\omega^{j-1})$.[5] By definition of precision of rounding $p_j^+ - p_j^- \ge \delta$ for all j.

[5] For technical reason, if necessary we add 0 and 1 to values of $f(\omega^{n-1})$ and set their probabilities be 0.

Denote $p_j = f(\omega^{j-1})$, $j = 1, 2, \ldots$. By definition p_j is a random variable. In the following we use the inequality

$$\phi_{t,\kappa_s}(\omega^{j-1}) \leq Pr\{p_j \geq 0.5\} \leq \phi_{t,\kappa_s}(\omega^{j-1}) + \kappa_s.$$

Consider two random variables

$$\vartheta_{n,1} = \sum_{j=k+1}^{n} \xi(p_j \geq 0.5)(\omega_j - p_j), \tag{10}$$

$$\vartheta_{n,2} = \sum_{j=k+1}^{n} \xi(p_j < 0.5)(\omega_j - p_j), \tag{11}$$

where $\xi(true) = 1$, and $\xi(false) = 0$.

We compute the bounds of mathematical expectations of these variables. These expectations are taken with respect to the overall probability distribution Pr generated by probability distributions Pr_j of random variables p_j, $j = 1, 2, \ldots$ (ω is fixed). Using the definition of the subword $\sigma \in G(i)$ of the sequence ω, we obtain ($k < j \leq n$)

$$E(\vartheta_{n,1}) \leq \sum_{\omega_j=0} Pr\{p_j \geq 0.5\}(-p_j^+) + \sum_{\omega_j=1} Pr\{p_j \geq 0.5\}(1 - p_j^+) \leq \tag{12}$$

$$-0.5 \sum_{j=k+1}^{n} \xi(\omega_j = 0)p_j^+ + (0.5 + \kappa_s) \sum_{j=k+1}^{n} \xi(\omega_j = 1)(1 - p_j^+).$$

$$E(\vartheta_{n,2}) \geq \sum_{\omega_j=0} Pr\{p_j < 0.5\}(-p_j^-) + \sum_{\omega_j=1} Pr\{p_j < 0.5\}(1 - p_j^-) \geq \tag{13}$$

$$-0.5 \sum_{j=k+1}^{n} \xi(\omega_j = 0)p_j^- + (0.5 - \kappa_s) \sum_{j=k+1}^{n} \xi(\omega_j = 1)(1 - p_j^-).$$

Subtracting (12) from (13) we obtain

$$E(\vartheta_{n,2}) - E(\vartheta_{n,1}) \geq 0.5 \sum_{j=k+1}^{n} \xi(\omega_j = 0)(p_j^+ - p_j^-) +$$

$$0.5 \sum_{j=k+1}^{n} \xi(\omega_j = 1)(p_j^+ - p_j^-) - \kappa_s \sum_{j=k+1}^{n} \xi(\omega_j = 1)(2 - p_j^- - p_j^+) \geq$$

$$\geq 0.5\delta(n-k) - 2\kappa_s(n-k) = (0.5\delta - 2\kappa_s)(n-k). \tag{14}$$

Then

$$E(\vartheta_{n,1}) \leq (-0.25\delta - \kappa_s)(n-k)$$

or

$$E(\vartheta_{n,2}) \geq (0.25\delta - \kappa_s)(n-k)$$

for infinitely many n, k. Since for any fixed f_t the ratio $k/n = i^{-1}$ and the number $\kappa_s = 1/s$ become arbitrary small for large i such that $i = \langle t, s\rangle$ for some s, we have

$$\liminf_{n\to\infty} \frac{1}{n} E(\vartheta_{n,1}) \le -0.25\delta$$

or

$$\limsup_{n\to\infty} \frac{1}{n} E(\vartheta_{n,2}) \ge 0.25\delta.$$

The martingale strong law of large numbers: for $\nu = 1, 2$, with Pr-probability one

$$\frac{1}{n}\sum_{j=1}^{n} I_\nu(p_j)(\omega_j - p_j) - \frac{1}{n} E(\vartheta_{n,\nu}) \to 0$$

as $n \to \infty$, implies that for $\nu = 0$ or for $\nu = 1$ the overall probability of the event (2) equals one. Lemma 6 and Theorem 1 are proved. □

Note, that inequalities (14) show that condition (2) of Theorem 1 can be replaced on

$$\limsup_{n\to\infty} \left| \frac{1}{n}\sum_{j=1}^{n} I_\nu(p_j)(\omega_j - p_j) \right| \ge 0.25\delta - \kappa$$

if in the construction of our algorithm the function (6) is computed up to a fixed precision κ.

The proof of Theorem 2 is in the line of the proof of Theorem 1, where $\phi(\omega^{n-1})$ denote a deterministic forecasting system. We have in the proof of Lemma 6, for some $\nu = 0$ or $\nu = 1$,

$$\sum_{j=k+1}^{n} I_\nu(p_j)(\omega_i - p_j) \ge (0.5 - 2\kappa_s)(n - k) \tag{15}$$

for infinitely many $k, n = ik$, where $p_j = f_i(\omega^{j-1})$, $j = 1, 2, \ldots$.

To prove (3) of Theorem 3 we define in (6) $\phi(\omega^{j-1}) = E(f(\omega^{j-1}))$ - the mathematical expectation of a random variable $f(\omega^{j-1})$. Then in the proof of Lemma 6, for some $\nu = 0$ or $\nu = 1$, the inequality (15), where p_j is replaced on $E(f(\omega^{j-1}))$, holds for infinitely many n. By the martingale strong law of large numbers we obtain that for $\nu = 0$ and for $\nu = 1$ with the overall probability one

$$\frac{1}{n}\sum_{j=1}^{n} I_\nu(E(p_j))(p_j - E(p_j)) \to 0 \tag{16}$$

as $n \to \infty$. Combining (16) with (15) modified as above, we obtain (3).

Acknowledgements

This research was partially supported by Russian foundation for fundamental research: 03-01-00475-a; 06-01-00122-a. A part of this work was done while the author was in Poncelet Laboratoire LIF CNRS, Marseille, France.

References

1. Dawid, A.P.: The well-calibrated Bayesian [with discussion]. J. Am. Statist. Assoc. 77, 605–613 (1982)
2. Dawid, A.P.: Calibration-based empirical probability [with discussion]. Ann. Statist. 13, 1251–1285 (1985)
3. Dawid, A.P.: The impossibility of inductive inference. J. Am. Statist. Assoc. 80, 340–341 (1985)
4. Foster, D.P., Vohra, R.: Asymptotic calibration. Biometrika 85, 379–390 (1998)
5. Kakade, S.M., Foster, D.P.: Deterministic calibration and Nash equilibrium. In: Shawe-Taylor, J., Singer, Y. (eds.) COLT 2004. LNCS (LNAI), vol. 3120, pp. 33–48. Springer, Heidelberg (2004)
6. Oakes, D.: Self-calibrating priors do not exists [with discussion]. J. Am. Statist. Assoc. 80, 339–342 (1985)
7. Rogers, H.: Theory of recursive functions and effective computability. McGraw-Hill, New York (1967)
8. Vovk, Vladimir, Takemura, Akimichi, Shafer, Glenn: Defensive Forecasting. In: Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. pp. 365–372 (2005), http://arxiv.org/abs/cs/0505083
9. V'yugin, V.V.: Non-stochastic infinite and finite sequences. Theor. Comp. Science. 207, 363–382 (1998)
10. Uspensky, V.A., Semenov, A.L., Shen, A.Kh.: Can an individual sequence of zeros and ones be random. Russian Math. Surveys 45(1), 121–189 (1990)
11. Zvonkin, A.K., Levin, L.A.: The complexity of finite objects and the algorithmic concepts of information and randomness. Russ. Math. Surv. 25, 83–124 (1970)