On Calibration Error of Randomized Forecasting Algorithms

Vladimir V. V'yugin [1]

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

Abstract

It has been recently shown that calibration with an error less than $\Delta > 0$ is almost surely guaranteed with a randomized forecasting algorithm, where forecasts are obtained by random rounding the deterministic forecasts up to Δ. We show that this error cannot be improved for a vast majority of sequences: we prove that, using a probabilistic algorithm, we can effectively generate with probability close to one a sequence "resistant" to any randomized rounding forecasting with an error much smaller than Δ. We also reformulate this result by means of a probabilistic game.

Key words: Machine learning, Universal prediction, Randomized prediction, Algorithmic prediction, Calibration, Randomized rounding

1 Introduction

A minimal requirement for testing any prediction algorithm is that it should be calibrated (see Dawid [1]). An informal explanation of calibration can be given as follows. Let a binary sequence $\omega_1, \omega_2, \dots, \omega_{n-1}$ of outcomes be observed by a forecaster whose task is to give a probability p_n of a future event $\omega_n = 1$. In a typical example 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. For simplicity, we consider binary sequences, i.e. $\omega_n \in \{0, 1\}$ for all n. We give a rigorous definition of calibration later.

[1] This research was partially supported by Russian foundation for fundamental research: 06-01-00122-a.

If the weather acts adversarially, then Oakes [11] and Dawid [2] show that any deterministic forecasting algorithm will not always be calibrated. V'yugin [20] presented a computable version of this result: he proved that no deterministic algorithm can be calibrated for a vast majority of sequences generated by some 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 internal randomization and the forecasts are hidden from the weather until weather makes its decision whether to rain or not.

Kakade and Foster [7] presented "an almost deterministic" *randomized rounding* universal forecasting algorithm f. For any sequence of outcomes and for any precision of rounding $\Delta > 0$ an observer can simply randomly round the deterministic forecast up to Δ in order to calibrate for this sequence with calibration error less than Δ.

The goal of this paper is to complement the result of Kakade and Foster with a lower bound. We prove in Theorems 1 and 2 that the upper bound Δ cannot be improved to $c\Delta$ for $c < 0.25$. We also show that when the setting of Theorem 2 is slightly modified calibration becomes impossible.

In Section 2 we discuss some approaches to universal prediction of individual sequences in statistics and machine learning. In particular, we give Dawid's definition of calibration and discuss its generalizations.

In Section 3 we give the definition of randomized forecasting and randomized rounding and formulate Kakade and Foster's result; the construction of the universal randomized rounding algorithm is given in Appendix A.

In Section 4 we consider some probabilistic version of the Oakes' example for randomized rounding algorithms. We present this example in the form of a game between Realty, Forecaster and Skeptic in the manner of Shafer and Vovk's book [14]. We show that Realty and Skeptic win with probability one if Forecaster randomly rounds deterministic forecasts up to a given positive precision level.

In Section 5 the asymptotic calibration is considered in the algorithmic setting. We present a uniform lower bound of calibration error for all computable randomized forecasting schemes. We also show that this uniform lower bound is valid for a vast majority of sequences generated by some probabilistic algorithm. The main result of this paper - Theorem 2, shows that given $\Delta > 0$ it is possible, using the probabilistic algorithm, to effectively generate with probability close to one a sequence "resistant" to any randomized rounding forecasting with an error much smaller than Δ. The proof of the main result is given in Section 6. Theorem 3 shows that when the setting of Theorem 2 is slightly modified the randomized forecasting with calibration error less than 0.25 becomes impossible. Theorem 4 shows that for a vast majority of sequences the calibration error can be much bigger, namely ≥ 0.25, irrespective of the precision of rounding we use when forecasts are produced by a deterministic forecasting algorithm.

This paper is an extended version of the conference paper [21].

2 Prediction of individual sequences

Predicting sequences is the key problem of machine learning and statistics. The learning process proceeds as follows: observing a finite-state sequence given on-line a forecaster assigns subjective probabilities to future states. The method of evaluation of these forecasts depends on an underlying learning approach.

According to the classical approach of statistics, we suppose that the observed sequence is generated by some source - a finite-state stochastic process governed by an unknown (to the forecaster) probability distribution. Asymptotic accuracy of forecasts is achieved when a forecast merges to the objective distribution. In the classical statistical theory of sequential prediction, the sequence of outcomes is assumed to be a realization of a stationary stochastic process. In this case statistical properties of the process based on past observations can be estimated and using this estimation efficient prediction strategies can be constructed. In this case the performance of a prediction strategy is usually evaluated by the expected value of some loss function which measures the distance between the predicted value and the true outcome (see, for example, [10]).

In the case of an arbitrary source distribution, the existence of a universal forecasting scheme was proved in the algorithmic information theory.

Solomonoff [16,17] and Levin [23] have defined a universal prior $M(x)$ as the probability that the output of a universal Turing machine starts with a sequence x when provided with fair coin flips on the input tape. Assume now that the sequences x are generated by a probability distribution P, i.e. the probability of a sequence starting from x is $P(x) > 0$. Then the probability of observing $\omega_n = 1$ after observations $\omega_1 \dots \omega_{n-1}$ is

$$P(1|\omega_1 \dots \omega_{n-1}) = P(\omega_1 \dots \omega_{n-1}1)/P(\omega_1 \dots \omega_{n-1}).$$

Solomonoff [17] proved that P-almost surely the universal posterior

$$M(1|\omega_1 \dots \omega_{n-1}) = M(\omega_1 \dots \omega_{n-1}1)/M(\omega_1 \dots \omega_{n-1})$$

converges to $P(1|\omega_1 \dots \omega_{n-1})$ as $n \to \infty$ if the measure P is computable. Solomonoff's result was further developed by Hutter [5] and Chernov et al. [6].

Hence, M is a valid predictor in the case of unknown computable P. Unfortunately, the function $M(1|\omega_1 \dots \omega_{n-1})$ is non-computable, so, it can be helpful only for a theoretical analysis.

According to a different viewpoint, we abandon the assumption that the outcomes are generated by a well-behaved stochastic process and consider the sequence of outcomes as produced by some unspecified mechanism, which could be deterministic, stochastic or adversarially adaptive to our prediction method. This setup where no probabilistic assumption is made on how the sequence is generated is often referred as prediction of individual sequences.

At the same time, without a probabilitic model it is not obvious how to measure the performance of the prediction algorithm. Some solution of this problem was proposed by Dawid [1], whose *prequential principle* says that our evaluation of the accuracy of the forecasts should not depend on any model of generating data. Dawid started from an observation that in reality, we have only individual sequence $\omega_1, \omega_2, \dots, \omega_n, \dots$ of outcomes and that the corresponding forecasts $p_1, p_2, \dots, p_n, \dots$ whose testing is considered may fall short of defining a full probability distribution on the whole set of all sequences. The prequential principle says that the evaluation of a probability forecaster should depend only on his actual probability forecasts and the corresponding outcomes.

The notion of calibration, originated by Dawid [1,2], checks whether the observed empirical frequencies of state occurrences converge to their forecaster probabilities. The notion of calibration complies with the prequential principle.

Let $I(p)$ denote the characteristic function of a subinterval $I \subseteq [0,1]$, i.e., $I(p) = 1$ if $p \in I$, and $I(p) = 0$, otherwise. An infinite sequence of forecasts $p_1, p_2, \dots$ is *well-calibrated* for an infinite sequence of outcomes $\omega_1\omega_2 \dots$ if for the characteristic function $I(p)$ of any subinterval of $[0,1]$ *the calibration error* tends to zero, i.e.,

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

as the denominator of the relation (1) tends to infinity.

The indicator function $I(p_i)$ determines some "selection rule" which selects indices i where we compute the deviation between forecasts p_i and outcomes ω_i. The most general notion of selection rule was considered in Vovk and

Shafer [18]: a selection rule is a function $F(p_1, \omega_1, \ldots, p_{i-1}, \omega_{i-1}, p_i)$ with range $\{0, 1\}$, where $\omega_i \in \{0, 1\}$ and $p_i \in [0, 1]$, $i = 1, 2, \ldots$.

The main problem of sequential forecasting is to define a universal forecasting algorithm which computes forecasts p_n given past observations $\omega_1, \ldots, \omega_{n-1}$ for each n. This universal prediction algorithm should be well-calibrated for each infinite sequence of outcomes. Oakes [11] proposed arguments (see Dawid [3] for a different proof) that no such algorithm can be well-calibrated for all possible sequences: any forecasting algorithm cannot be 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 are forecasts computed by the algorithm given $\omega_1, \ldots, \omega_{i-1}$, $i = 1, 2, \ldots$. The corresponding intervals are $I_0 = [0, 0.5)$ and $I_1 = [0.5, 1]$. It is easy to see that the condition (1) of calibration fails for this ω, where $I = I_0$ or $I = I_1$.

Theorem 4 given below in Section 5 presents an effective method for generating a vast majority of sequences possessing this property. It says that using coin-tossing and a transducer algorithm, we can generate with probability close to one an infinite sequence ω such that each forecasting algorithm cannot be calibrated for ω.

Foster and Vohra [4] show that calibration is almost surely guaranteed with a randomizing forecasting rule, i.e., where the forecasts are chosen using internal randomization. Kakade and Foster [7] noticed that some calibration results require very little randomization. They defined "an almost deterministic" *randomized rounding* universal forecasting algorithm f: an observer can only randomly round the deterministic forecast in order to calibrate for each sequence of outcomes regardless of the nature of the source generating it. This approach was further developed by, among others, Lehrer [8], Sandrony et al. [13]. These papers were only concerned with asymptotic calibration. These authors asked only that that the entire sequence of forecasts and its certain subsequences be properly calibrated.

Non-asymptotic version of randomized forecasting was proposed by Vovk and Shafer [18] and by Vovk et al. [19]. This approach is based on the game-theoretic framework of Shafer and Vovk [14]. The main requirement of this approach that the forecasts resist any betting strategy can be interpreted by saying that they must pass all statistical tests, not only tests of calibration.

We discuss details of the randomized forecasting algorithms in Section 3 and in Appendix A. The game-theoretic framework is partially used in Section 4.

In this section we also develop a probabilistic version of the Oakes' example for the randomized rounding algorithms and present the corresponding betting strategy.

3 Randomized forecasting

Forecasting can be thought as a perfect-information game between two players: Forecaster and Realty [18]. Assuming that Realty makes a binary choice at each step, *a game of deterministic forecasting* is described by the protocol:

FOR $n = 1, 2, \ldots$
Forecaster announces a forecast $p_n \in [0,1]$.
Reality announces an outcome $\omega_n \in \{0,1\}$.
ENDFOR

Let $\mathcal{P}[0,1]$ be the set of all probability measures on the unit interval $[0,1]$ supplied with the standard σ-algebra $\mathcal{F}$ of all measurable subsets of $[0,1]$.

A game of randomized forecasting requires a new player - Random Number Generator. It can be described by the protocol:

FOR $n = 1, 2, \ldots$
Forecaster announces probability distribution $P_n \in \mathcal{P}[0,1]$.
Reality announces $\omega_n \in \{0,1\}$.
Random Number Generator announces $p_n \in [0,1]$ distributed according to P_n.
ENDFOR

By Sandrony et al. [13] *a randomized forecasting scheme* is a sequence of functions $\zeta_n : \{0,1\}^{n-1} \times [0,1]^{n-1} \to \mathcal{P}[0,1]$; for any n given a sequence of past forecasts $p_1, \ldots, p_{n-1}$ and a sequence of outcomes $\omega^{n-1} = \omega_1 \ldots \omega_{n-1}$, ζ_n outputs a probability P_n for a forecast p_n. We denote the randomized forecasting scheme $P_{\omega^{n-1}}(\cdot|p_1, \ldots, p_{n-1})$, $n = 1, 2, \ldots$ ($\omega^0 = \lambda$ is the empty sequence).

For any such randomized forecasting scheme by Ionescu-Tulcea theorem (see Shiryaev [15]) for any sequence of outcomes $\omega = \omega_1, \omega_2, \ldots$, a unique overall probability measure Pr on the set $[0,1]^\infty$ of all forecasting paths - infinite sequences $p_1, p_2, \ldots$, exists such that for each n, $P_{\omega^{n-1}}(\cdot|p_1, \ldots, p_{n-1})$ is a conditional probability induced by Pr, i.e.,

$$P_{\omega^{n-1}}(p_n \in A | p_1, \ldots, p_{n-1}) = Pr(p_n \in A | p_1, \ldots, p_{n-1})$$

for all n, where A is a measurable subset of $[0,1]$.

Assume that for each n, the probability distribution $P_{\omega^{n-1}}(\cdot|p_1,\dots,p_{n-1})$ is concentrated on a finite subset D_n of $[0,1]$, say, $D_n=\{p_{n,1},\dots,p_{n,m_n}\}$. The number $\Delta=\liminf\limits_{n\to\infty}\Delta_n$, where

$$\Delta_n=\inf\{|p_{n,i}-p_{n,j}|:i\ne j\},$$

is called *the level of discreteness* of the corresponding forecasting scheme on the sequence $\omega=\omega_1\omega_2\dots$.

In general case D_n is a predictable, i.e., measurable with respect to the σ-algebra $\mathcal{F}^{n-1}$, random variable depending on ω^{n-1}.

A typical example is the uniform rounding: for each n, rational points $p_{n,i}$ divide the unit interval into equal parts of size $0<\Delta<1$; the probability distribution $P_{\omega^{n-1}}(\cdot|p_1,\dots,p_{n-1})$ is concentrated on these points. In this case the corresponding level of discreteness on arbitrary sequence $\omega_1,\omega_2,\dots$ equals Δ.

Kakade and Foster [7] presented "an almost deterministic" universal randomized rounding forecasting scheme. This scheme, given a partition of the unit interval into equal parts of size Δ, randomly rounds a deterministic forecast computed by some algorithm to nearby points of the partition.

Notice that the random forecast p_n generated by Kakade and Foster's forecasting scheme is independent on the previous forecasts $p_1,\dots,p_{n-1}$; it depends only on ω^{n-1} (see Appendix A). In this case the overall probability distribution Pr defined by Kakade and Foster's randomized forecasting scheme $P_{\omega^{n-1}}$ is the product of independent probability distributions.

Proposition 1 *(Kakade and Foster) For any infinite sequence $\omega=\omega_1\omega_2\dots$ and for the characteristic function $I(p)$ of any subinterval of $[0,1]$ the overall probability Pr of the event*

$$\left|\frac{1}{n}\sum_{i=1}^{n}I(p_i)(\omega_i-p_i)\right|\le\Delta \tag{2}$$

tends to 1 as $n\to\infty$, where Pr is the overall probability distribution defined by Kakade and Foster's randomized forecasting scheme $P_{\omega^{n-1}}$, $n=1,2,\dots$, and $p_1,p_2,\dots$ is a sequence of forecasts distributed according to Pr.

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 0.2688 can be rounded up to second digit to 0.26 with probability 0.12, and to 0.27 with probability 0.88.

Details of the Kakade and Foster's algorithm are given in Appendix A.

4 Oakes' example for randomized rounding forecasting

In this section we consider some probabilistic variant of the Oakes' example in the form of a game. This game will be used in the proof of Theorem 2.

A generalization of Foster and Vohra's result was obtained by Vovk and Shafer [18]. They consider a perfect-informatin game of randomized forecasting between the players - Forecaster, Skeptic, Reality, Random Number Generator.

The goal of Forecaster is to state probabilities that pass all possible statistical tests in light of Realty subsequent moves. This goal is formalized by adding a third player - Skeptic, who seeks to refute Forecaster's probabilities.

We consider some probabilistic version of the game:

Let $\mathcal{K}_0 = 1$.
FOR $n = 1, 2, \dots$
Skeptic announced $S_n : [0, 1] \to \mathcal{R}$.
Forecaster announces a probability distribution $P_n \in \mathcal{P}[0, 1]$.
Reality announces $\omega_n \in \{0, 1\}$.
Random Number Generator announces $p_n \in [0, 1]$ distributed according to P_n.
Skeptic updates his capital $\mathcal{K}_n = \mathcal{K}_{n-1} + S_n(p_n)(\omega_n - p_n)$.
ENDFOR

Restriction on Skeptic: Skeptic must choose the S_n so that his capital $\mathcal{K}_n$ is nonnegative for all n no matter how the other players move.

Winner: Forecaster wins if Skeptic's capital $\mathcal{K}_n$ stays bounded as $n \to \infty$. Otherwise Realty and Skeptic win. Random Number Generator is a *neutral* player. [2]

Theorem 1 of Vovk and Shafer [18] which is stated in a purely game-theoretic setting implies for our mixed setting that there exist a randomized strategy for Forecaster - a sequence of probability distributions
$P_n = P_{\omega^{n-1}}(\cdot|p_1, \dots p_{n-1})$, $n = 1, 2, \dots$, and an overall probability Pr for these probabilities such that for each sequence $\omega = \omega_1\omega_2\dots$ announced by Realty the set of forecasting paths $p_1, p_2, \dots$ where Forecaster wins has Pr-probability 1.

[2] Our approach is not purely game-theoretic in sense of Shafer and Vovk [14]. Random Number Generator is used to introduce into consideration the forecasting paths. A purely game-theoretic version of this game and of Theorem 1 (and even of Theorem 2) can also be obtained, but it is out of the scope of this paper.

By [18] (Theorem 3) and [19] Proposition 1 is a corollary of this result.

We prove that when Forecaster uses finite subsets of $[0,1]$ for randomization Realty and Skeptic can defeat Forecaster in this forecasting game.

Define a strategy for Realty: at step n Realty announces an outcome

$$\omega_n = \begin{cases} 0 \text{ if } P_n([0.5,1]) \ge 0.5 \\ 1 \text{ otherwise.} \end{cases}$$

Let $\epsilon_k = 2^{-k}$, $k = 1, 2, \ldots$. We consider two infinite sequences of Skeptic's auxiliary strategies:

$$S_n^{1,k}(p) = -\epsilon_k \mathcal{K}_{n-1}^{1,k} \xi(p \ge 0.5), \tag{3}$$

$$S_n^{2,k}(p) = \epsilon_k \mathcal{K}_{n-1}^{2,k} \xi(p < 0.5), \tag{4}$$

where $\xi(true) = 1$, $\xi(false) = 0$ and $\mathcal{K}_{n-1}^{s,k}$ is the capital of Skeptic at the end of step $n-1$ when he follows the strategy $S_{n-1}^{s,k}$, $s = 1, 2$ and $k = 1, 2, \ldots$, $\mathcal{K}_0^{s,k} = 1$.

We combine them in the strategy $S_n(p) = \frac{1}{2}(S_n^1(p) + S_n^2(p))$, where $S_n^1(p) = \sum\limits_{k=1}^{\infty} \epsilon_k S_n^{1,k}(p)$ and $S_n^2(p) = \sum\limits_{k=1}^{\infty} \epsilon_k S_n^{2,k}(p)$. It follows from the definition that $K_n^{s,k}(p) \le 2^n$ and $|S_n^{s,k}(p)| \le 2^{n-1}$ for $s = 1, 2$ and for all p, k and n. Then these sums are finite for each n and p.

The total capital of Skeptic at step n when he follows the strategy $S_n(p)$ equals

$$\mathcal{K}_n = \frac{1}{2} \sum_{k=1}^{\infty} \epsilon_k (\mathcal{K}_n^{1,k} + \mathcal{K}_n^{2,k}).$$

Suppose Forecaster uses a randomized forecasting scheme $P_{\omega^{n-1}}(\cdot|p_1, \ldots, p_{n-1})$ for defining P_n. Since ω^{n-1} and S_n^i, $i = 1, 2$, are defined in terms of σ-algebra $\mathcal{F}^{n-1}$, by Ionescu-Tulcea theorem an overall probability distribution Pr exists such that for each n, $P_n = Pr(\cdot|p_1, \ldots, p_{n-1})$ is the conditional probability induced by Pr.

Theorem 1 *Suppose Forecaster's randomized forecasting scheme has a positive level of discreteness on each infinite sequence ω. Then using the strategies defined above, Realty and Skeptic win with Pr-probability 1.*

Proof. Suppose that for each j, p_j is distributed according to P_j. Define two sequences of random variables

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

where $n = 1, 2, \ldots$.

Suppose that for any n, the probability distribution P_n is concentrated on a finite set $\{p_{n,1}, \ldots, p_{n,m_n}\}$. Denote $p_n^- = \max\{p_{n,t} : p_{n,t} < 0.5\}$ and $p_n^+ = \min\{p_{n,t} : p_{n,t} \geq 0.5\}$. [3] By definition ω_n, p_n^+ and p_n^- are predictable and $p_n^+ - p_n^- \geq \Delta$ for all n, where $\Delta > 0$. We have

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

where E is the mathematical expectation with respect to Pr. Then

$$E(\vartheta_{n,2}) - E(\vartheta_{n,1}) \geq 0.5\Delta n \tag{5}$$

for all n.

By the martingale strong law of large numbers with Pr-probability 1,

$$\frac{1}{n}\vartheta_{n,1} - \frac{1}{n}E(\vartheta_{n,1}) \to 0,$$
$$\frac{1}{n}\vartheta_{n,2} - \frac{1}{n}E(\vartheta_{n,2}) \to 0 \tag{6}$$

as $n \to \infty$. Then by (5) and (6) with Pr-probability 1 for each $\delta > 0$,

$$\frac{1}{n}\vartheta_{n,2} - \frac{1}{n}\vartheta_{n,1} \geq 0.5\Delta - \delta \tag{7}$$

for all sufficiently large n. In particular, (7) implies Corollary 1 below.

[3] For technical reason, if necessary, we add 0 and 1 to the support set of P_n and set their probabilities to be 0.

We follow Shafer and Vovk's [14] method of defining the defensive strategy for Skeptic. For any k and n,

$$\mathcal{K}_n^{1,k} = \prod_{j=1}^{n} (1 - \epsilon_k \xi(p_j \geq 0.5)(\omega_j - p_j)), \tag{8}$$

$$\mathcal{K}_n^{2,k} = \prod_{j=1}^{n} (1 + \epsilon_k \xi(p_j < 0.5)(\omega_j - p_j)). \tag{9}$$

If Skeptic follows the strategy (3) then at step n

$$\ln \mathcal{K}_n^{1,k} \geq -\epsilon_k \vartheta_{n,1} - \epsilon_k^2 n. \tag{10}$$

If Skeptic follows the strategy (4) then at step n

$$\ln \mathcal{K}_n^{2,k} \geq \epsilon_k \vartheta_{n,2} - \epsilon_k^2 n. \tag{11}$$

Here we have used the inequality $\ln(1+r) \geq r - r^2$ for all $|r| \leq 1$.

The inequalities (7), (10) and (11) imply

$$\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n^{1,k} + \ln \mathcal{K}_n^{2,k}}{n} \geq \frac{1}{2}\epsilon_k \Delta - 2\epsilon_k^2 \geq 2\epsilon_k^2 \tag{12}$$

is valid with Pr-probability 1, where $\epsilon_k \leq \frac{1}{8}\Delta$.

Using (12), for $s = 1$ or for $s = 2$ with Pr-probability 1

$$\limsup_{n\to\infty} \frac{\ln \mathcal{K}_n^{s,k}}{n} \geq \epsilon_k^2. \tag{13}$$

By definition for $s = 1, 2$ and for all k, $\mathcal{K}_n^{s,k} \geq 0$ for all n no matter how the other players move. By (13) if $\epsilon_k \leq \frac{1}{8}\Delta$ then for $s = 1$ or for $s = 2$, $\limsup_{n\to\infty} \mathcal{K}_n^{s,k} = \infty$ with Pr-probability 1.

Since $\mathcal{K}_n$ is the weighted sum of $\mathcal{K}_n^{s,k}$, we have $\limsup_{n\to\infty} \mathcal{K}_n = \infty$ on a set of forecasting paths $p_1, p_2, \ldots$ of Pr-probability 1. $\triangle$

We also obtain a lower bound of calibration error.

Corollary 1 *Suppose Forecaster's randomized forecasting scheme has a positive level of discreteness on each infinite sequence ω. Then using the strategy* *defined above, Realty announces a sequence ω such that, for $i = 1$ or for $i = 2$, with Pr-probability* 1

$$\limsup_{n\to\infty} \left| \frac{1}{n} \vartheta_{n,i} \right| \geq 0.25\Delta. \tag{14}$$

This corollary follows from (7).

5 Uniform lower bounds of calibration error

In this section we study the asymptotic calibration in the algorithmic setting. In particular, we present a computable uniform version of Theorem 1. We show that the lower bound (14) of calibration error also is a uniform lower bound for all computable randomized forecasting schemes. We also show that this uniform lower bound is valid for a vast majority of sequences generated by some probabilistic algorithm.

Let $\Omega = \{0,1\}^\infty$ be the set of all infinite binary sequences, $\Xi = \cup_{n=1}^\infty \{0,1\}^n$ be the set of all finite binary sequences, and λ be the empty sequence. For any finite or infinite sequence $\omega = \omega_1 \dots \omega_n \dots$, we write $\omega^n = \omega_1 \dots \omega_n$ (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.

We need some computability concepts. Let $\mathcal{R}$ be the set of all real numbers extended by adding the infinities $-\infty$ and $+\infty$.

For any set of of finite objects A, we suppose that elements of A can be effectively enumerated by positive integer numbers (see Rogers [12]). In particular, we will identify a computer program with 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) with its number $\langle t,s \rangle$.

A function $\phi: A \to \mathcal{R}$ is called (lower) semicomputable if the set

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

is recursively enumerable. 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 $\phi(x) \leq r$ we may be for ever uncertain. A function ϕ is upper semicomputable if $-\phi$ is lower semicomputable.

A standard argument based on the recursion theory shows that there exist the lower and the upper semicomputable real functions $\phi^-(j,x)$ and $\phi^+(k,x)$ universal for all lower semicomputable and upper semicomputable functions from $x \in \Xi$. [4] 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 enumeration 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.

A function $\phi : \Xi \to \mathcal{R}$ is computable if there exists an algorithm which given $z \in \Xi$ and a degree of precision κ computes a rational approximation of $\phi(z)$ up to κ. In more detail, let $i = \langle t,k\rangle$. We say that the function $\phi_i(x)$ is *defined on* x if given any degree of precision - a positive rational number $\kappa > 0$, it holds that

$$|\phi_s^+(t,x) - \phi_s^-(k,x)| \le \kappa \tag{15}$$

for some s; $\phi_i(x)$ is undefined otherwise. If some s exists such that (15) holds, define $\phi_{i,\kappa}(x) = \phi_s^-(k,x)$ for minimal such s, $\phi_{i,\kappa}(x)$ is undefined otherwise. The function $\phi_{i,\kappa}(x)$ is called the rational approximation (from below) of $\phi_i(x)$ up to κ.

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

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

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

We also use the concept of *computable operation* on $\Xi\bigcup\Omega$ [22,23]. Let $\hat{F}$ be a recursively enumerable set of ordered pairs of finite sequences satisfying the

[4] 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.

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 that follows $P = L$, where $L(x) = L(\Gamma_x) = 2^{-l(x)}$ is the uniform measure on Ω.

A deterministic forecasting system is a function $f : \Xi \to [0, 1]$ (see Dawid [1]). In that follows we consider total forecasting systems f, i.e., everywhere defined on Ξ.

In that follows, we mean by *a randomized forecasting system* a probability distribution P_x on $[0, 1]$, where $x \in \Xi$ is a parameter. [5] The precise definition of computable probability distribution on $[0, 1]$ requires some technicalities. In fact, in the construction below, we compute P_x only at one set $[0.5, 1]$; so, we call a randomized forecasting system P_x *weakly computable* if $P_x([0.5, 1])$ is a computable function from the parameter x.

Let $I_0 = I_0(p)$ and $I_1 = I_1(p)$ be the characteristic functions of the intervals $[0, 0.5)$ and $[0.5, 1]$, correspondingly.

The following theorem gives a uniform lower bound of calibration error for computable randomized forecasting with positive level of discretness.

[5] Such forecasting systems are used in Kakade and Foster's universal forecasting algorithm (see Appendix A). We can prove a game-theoretic version of Theorem 2 based on the more general Sandrony's et al. forecasting schemes.

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 $\omega = \omega_1\omega_2\dots$ such that for every weakly computable randomized forecasting system P_x with the level of discreteness Δ on ω, for $\nu = 0$ or for $\nu = 1$, Pr-probability of the event*

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I_\nu(p_i)(\omega_i - p_i)\right| \geq 0.25\Delta \tag{16}$$

equals 1, where the overall probability Pr is defined by $P_{\omega^{i-1}}$, $i = 1, 2, \dots$, and $p_1, p_2, \dots$ are distributed according to these probabilities.

Theorem 2 uses forecast-based selection rules - $I_\nu(p_i)$, $\nu = 0, 1$. In the case when we select random forecasts for checking using their mean value, we obtain a lower bound does not depending on the accuracy of randomized rounding.

Let for any i, $E_{\omega^{i-1}} = \int\limits_0^1 pP_{\omega^{i-1}}(dp)$ be the mean value of the forecasts generated by a randomized forecasting system $P_{\omega^{i-1}}$.

In the following theorem we consider randomized forecasting systems f with computable mathematical expectations, i.e., $E_{\omega^{i-1}}$ is a computable real function from the parameter ω^{i-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 $\omega = \omega_1\omega_2\dots$ such that for every randomized forecasting system P_x with computable mathematical expectation, for $\nu = 0$ or for $\nu = 1$, Pr-probability of the event*

$$\limsup_{n\to\infty}\left|\frac{1}{n}\sum_{i=1}^{n} I_\nu(E_{\omega^{i-1},n})(\omega_i - p_i)\right| \geq 0.25 \tag{17}$$

equals 1, where the overall probability Pr is defined by $P_{\omega^{n-1}}$, $n = 1, 2, \dots$, and $p_1, p_2, \dots$ are distributed according to these probabilities. [6] *Here, $E_{\omega^{i-1},n}$ is a computable sequence of rational approximations of $E_{\omega^{i-1}}$ up to a given precision κ_n, where $\kappa_n \to 0$ as $n \to \infty$, $i = 1, 2, \dots$.* [7]

The following theorem is a computable uniform version of the Oakes' counter example for deterministic forecasting systems. We expand this example to a

[6] The lower bounds of Theorems 3 and 4 have been weakened comparing to the conference version [21] (0.25 is used instead of $0.5 - \epsilon$). Author does not know how to improve these lower bounds.

[7] This sequence will be constructed in the proofs of Theorems 2 and 3.

large set of sequences generated by a probabilistic algorithm. A variant of this result was obtained by V'yugin [20].

An outcome-based selection rule is a function $\delta : \Xi \to \{0, 1\}$.

Theorem 4 *For any $\epsilon > 0$, a probabilistic algorithm (L, F) can be constructed, which with probability $\geq 1 - \epsilon$ outputs an infinite binary sequence $\omega_1\omega_2\ldots$ such that for every deterministic forecasting algorithm f there exists a computable outcome-based selection rule δ such that the following inequality holds*

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=1}^{n} \delta(\omega^{i-1})(\omega_i - p_i) \right| \geq 0.25, \tag{18}$$

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

For each deterministic forecasting system f, there exists a unique probability measure P on Ω such that $P(\Gamma_{x1}) = f(x)P(\Gamma_x)$ for all $x \in \Xi$. In other words, f is a version of the conditional probability, according to P, that x will be followed by 1. Dawid's original notion of calibration can be considered as a version of von Mises' definition of an individual random sequence with respect to the measure P (see Li and Vitányi [9] for details of algorithmic randomness).

V'yugin [20] proved that for any computable probability measure P, the martingale strong law of large numbers holds for each infinite sequence ω random with respect to P in sense of Martin-Löf. Therefore, the probabilistic algorithm (L, F) from Theorem 4 outputs with probability $\geq 1 - \epsilon$ an infinite binary sequence ω which is not Martin-Löf random (it is not even Dawid random) with respect to each computable probability distribution.

6 Proofs of Theorems 2-4

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

$$Q(x) = P\{\omega : x \sqsubseteq F(\omega)\}. \tag{19}$$

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

$$\begin{aligned} Q(\lambda) &\leq 1; \\ Q(x0) + Q(x1) &\leq Q(x) \end{aligned}$$

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 (19) holds, where $P = L$ (for the proof see [22,23]).

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

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

for all $x \in \Xi$ and extend this function to all Borel subsets of Ω (see [23]).

For any semimeasure P, define *the support set of* P

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

It is easy to see that E_P is a closed subset of Ω and $\bar{P}(E_P) = \bar{P}(\Omega)$.

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 directed edges $(x, x0)$ and $(x, x1)$ of length one. In the construction below we will add 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 an 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 every $x \in \Xi$

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

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

$$R(\lambda) = 1;$$
$$R(y) = \sum_{\sigma:\sigma_2 = y} q(\sigma) R(\sigma_1) \tag{20}$$

for $y \neq \lambda$.

By definition R is a computable function taking rational values and the semimeasure P is lower semicomputable.

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 programs i computing the functions $\phi_i(x)$. [8] In the proof (see Lemma 6) we use a special class of these functions, namely, functions of the type

$$\phi(x) = P_x([0.5, 1]), \tag{21}$$

where P_x, $x \in \Xi$, is a weakly computable randomized forecasting system. For any such function ϕ, $\phi = \phi_i$ for some i.

The processing of any program i is divided on sessions $s = 1, 2, \ldots$. At each session s, the construction works with the rational approximations $\phi_{i,\kappa_s}(x)$ of $\phi_i(x)$ from below up to κ_s for all $x \in \Xi$, where $\kappa_s = 1/s$.

Besides, at each session s, we visit each rational approximation $\phi_{i,\kappa_s}(x)$ on infinitely many steps n. To do this, we define some function $p(n)$ such that for any positive integer number m we have $p(n) = m$ for infinitely many n. For example, we can define $p(\langle m, k\rangle) = m$ and $p'(\langle m, k\rangle) = k$ for all m and k, where $\langle m, k\rangle$ is some computable one-to-one enumeration of all pairs of nonnegative integer numbers. Then for each step n we compute $\langle i, s\rangle = p(n)$, where i is a program and s is a number of a session; so, $i = p(p(n))$ and $s = p'(p(n))$.

For a program i, a number of session s, finite binary sequences x and y, an elementary network q, and for a nonnegative integer number n, the predicate $B(\langle i, s\rangle, x, y, q, n)$ is *true* if the following holds

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

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

[8] Recall that $i = \langle j, k\rangle$ 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 $\phi_i(x)$.

It is *false* otherwise. Define

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

Here $p(p(l(x)))$ is a program and $p'(p(l(x)))$ is a number of a session; min is considered for lexicographical ordering of strings; we suppose that $\min\emptyset$ is undefined.

Lemma 1 *For any total function ϕ_i, $\beta(x,q,n)$ is defined for all $x \in \Xi$ and for all sufficiently large n such that $p(p(n)) = i$.*

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

The goal of the construction below is the following. Each extra edge σ corresponds to some task number $I = \langle i, s\rangle$ 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 each infinite binary sequence ω the following holds: 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 each extra edge σ added to the network q, $B(I, \sigma_1, \sigma_2, q^{n-1}, n)$ is true; it is false, otherwise.

Lemma 5 shows that $\bar{Q}(E_Q) > 1 - 0.5\epsilon$, where Q is the q-flow and E_Q is its support set. Lemma 6 shows that for each $\omega \in E_Q$, the event (16) holds with the overall probability 1.

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} be 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 also suppose 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 three cases.

Let $w(I, q^{n-1})$ be equal to the minimal m such that $p(m) = I$ and $m > sl(\sigma_2)$ for each extra edge $\sigma \in G^{n-1}$ such that $p(l(\sigma_1))) < I$, where $s = p'(p(I))$ is the number of the session assigned with the task I.

The inequality $w(I, q^m) \neq w(I, q^{m-1})$ can be induced by some task $J < I$

that adds 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 $G^n = G^{n-1}$.

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 [9] 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 $I = p(n)$ *adds* 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 $\sigma \in G^n$.

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 $q = \lim_{n\to\infty} q^n$, the network flow delay function $d = \lim_{n\to\infty} d^n$ and the set of extra edges $G = \cup_n G^n$.

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

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

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

Lemma 2 *The set $G(I)$ is finite 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 all I. Suppose 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 $w(I,q) < \infty$.

By definition if $d(\omega^m) \neq 0$ then $p_m = 1/d(\omega^m)$ is a positive integer number. Besides, if $(\omega^n, y), (\omega^m, y') \in G(I)$, where $n < m$ and $l(y) = m$, then $p_n > p_m$. Hence, for each $\omega \in \Omega$ a maximal m exists such that $(\omega^m, y) \in G(I)$ for some y or no such extra edge exists. In the latter case let $m = w(I,q)$. Define $u(\omega) = 1/d(\omega^m)$.

By the construction the integer valued function $u(\omega)$ is constant on the interval Γ_{ω^m}, and then, it is continuous in the topology generated by such intervals. Since Ω is compact in this topology, $u(\omega)$ is bounded. Then for some m', $u(\omega) = u(\omega^{m'})$ for all ω. By the construction if any extra edge of Ith type was added to $G(I)$ at some step then $d(y) > d(x)$ holds for some new pair (x,y) such that $x \subseteq y$. This gives us a contradiction if $G(I)$ is infinite. $\triangle$

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 an extra edge then $B(I, \sigma_1, \sigma_2, n)$ is true, where $n = l(\sigma_2)$.

Lemma 3 *Assume that for each 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. Assume 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 added 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 choice of m we have $y \sqsubseteq \omega$. $\triangle$

Lemma 4 *For any y, $Q(y) = 0$ 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, assume 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 cannot be added to the network $q^{l(z)-1}$ since $d(z^n) = 1$. This contradiction proves the lemma. $\triangle$

By Lemma 4 the relation $Q(y) = 0$ is recursive and

$$E_Q = \Omega \setminus \cup_{d(x)=1} \Gamma_x, \tag{22}$$

where E_Q is the support set of Q.

Lemma 5 *It holds that $\bar{Q}(E_Q) > 1 - 0.5\epsilon$.*

Proof. We bound $\bar{Q}(\Omega)$ from below. Let R be defined by (20). By definition of the network flow delay function, we have

$$\sum_{u:l(u)=n+1} R(u) = \sum_{u:l(u)=n} (1 - d(u))R(u) + \sum_{\sigma:\sigma\in G, l(\sigma_2)=n+1} q(\sigma)R(\sigma_1). \tag{23}$$

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 $S_{n+1} \geq S_n$. Assume 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

$$\begin{aligned} \sum_{u:l(u)=n} d(u)R(u) &= \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} d(\sigma_2) \sum_{u:P(u,\sigma)} R(u) = \\ = \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} \frac{d(\sigma_1)}{1 - d(\sigma_1)} \sum_{u:P(u,\sigma)} R(u) &\leq \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} d(\sigma_1)R(\sigma_1) = \\ &= \sum_{\sigma:\sigma\in G, l(\sigma_2)=n} q(\sigma)R(\sigma_1). \end{aligned} \tag{24}$$

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 (23) 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 (23) and (24) we obtain

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

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

$$S_n \ge 1 - \sum_{i=1}^{\infty} \frac{1}{(i+n_0)^2} \ge 1 - 0.5\epsilon$$

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

$$\bar{Q}(\Omega) = \inf_n \sum_{l(u)=n} Q(u) \ge \inf_n S_n \ge 1 - 0.5\epsilon.$$

Lemma is proved. $\triangle$

Lemma 6 *For each weakly computable randomized forecasting system P_x and and for each sequence $\omega \in E_Q$, the event (16) holds with Pr-probability* 1, *where the overall probability Pr is defined by $P_{\omega^{i-1}}$, $i = 1, 2, \dots$.*

Proof. Assume that $\omega \in E_Q$ and P_x is a weakly computable randomized forecasting system, i.e., the corresponding $\phi_i(x) = P_x([0.5, 1])$ is defined for all $x \in \Xi$.

Let ϕ_{i,κ_s} be a rational approximation of of ϕ_i from below up to $\kappa_s = 1/s$, and let $I = \langle i, s \rangle$. Since there are infinitely many sessions s of the construction when we visit ϕ_i, we can consider only steps n, $p(n) = I$, such that s is sufficiently large.

By (22), $d(\omega^n) < 1$ for all n. 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 that follows $k = l(\sigma_1)$ and $n = sk$.

We reproduce the game from the proof of Theorem 1 on the edge σ.

Recall that for any j, $p_j^- = \max\{p_{j,t} : p_{j,t} < 0.5\}$ and $p_j^+ = \min\{p_{j,t} : p_{j,t} \geq 0.5\}$, where $\{p_{j,1}, \dots, p_{j,m_j}\}$ is the support set of $P_{\omega^{j-1}}$. By definition of precision of rounding $p_j^+ - p_j^- \geq \Delta$ for all j.

Assume that p_j are distributed according to $P_{\omega^{j-1}}$, where $j = 1, 2, \dots$. In that follows, we use the inequality

$$\phi_{i,\kappa_s}(\omega^{j-1}) \leq P_{\omega^{j-1}}\{p_j \geq 0.5\} \leq \phi_{i,\kappa_s}(\omega^{j-1}) + \kappa_s. \tag{25}$$

Consider two sequences of random variables

$$\vartheta_{n,1} = \sum_{j=k+1}^{n} \xi(p_j \geq 0.5)(\omega_j - p_j), \tag{26}$$

$$\vartheta_{n,2} = \sum_{j=k+1}^{n} \xi(p_j < 0.5)(\omega_j - p_j), \tag{27}$$

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

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

$$\begin{aligned} E(\vartheta_{n,1}) \leq \sum_{\omega_j=0} P_{\omega^{j-1}}\{p_j \geq 0.5\}(-p_j^+) + \\ \sum_{\omega_j=1} P_{\omega^{j-1}}\{p_j \geq 0.5\}(1 - p_j^+) \leq \\ -0.5 \sum_{j=k+1}^{n} \xi(\omega_j = 0)p_j^+ + (0.5 + \kappa_s) \sum_{j=k+1}^{n} \xi(\omega_j = 1)(1 - p_j^+). \end{aligned} \tag{28}$$

$$\begin{aligned} E(\vartheta_{n,2}) \geq \sum_{\omega_j=0} P\omega^{j-1}\{p_j < 0.5\}(-p_j^-) + \\ \sum_{\omega_j=1} P_{\omega^{j-1}}\{p_j < 0.5\}(1 - p_j^-) \geq \\ -0.5 \sum_{j=k+1}^{n} \xi(\omega_j = 0)p_j^- + (0.5 - \kappa_s) \sum_{j=k+1}^{n} \xi(\omega_j = 1)(1 - p_j^-). \end{aligned} \tag{29}$$

Subtracting (28) from (29) we obtain

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

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

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

$$\ge 0.5\Delta(n-k)-2\kappa_s(n-k)=(0.5\Delta-2\kappa_s)(n-k). \quad (30)$$

Then

$$E(\vartheta_{n,1})\le(-0.25\Delta-\kappa_s)(n-k)$$

or

$$E(\vartheta_{n,2})\ge(0.25\Delta-\kappa_s)(n-k)$$

for infinitely many n,k. Since the ratio k/n and the number $\kappa_s=1/s$ become arbitrary small for large n, we have

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

or

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

By the strong law of large numbers, for $\nu=1,2$, with Pr-probability 1

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

as $n\to\infty$. This implies that Pr-probability of the event (16) equals 1 for $\nu=0$ or for $\nu=1$. Lemma 6 and Theorem 2 are proved. $\triangle$

Sketch of the proof of Theorem 4. The proof is in the line of the proof of Theorem 2, where $\phi(\omega^{n-1})$ denotes a deterministic forecasting system. Let n, k be the same as in the proof of Lemma 6, and let κ_s be a non-increasing computable function taking rational values such that $\kappa_s\to 0$ as $s\to\infty$. By definition of the sequence ω for $\nu=0$ or for $\nu=1$, $I_\nu(\phi_{i,\kappa_s}(\omega^{j-1}))=1$ for at least one half of all j such that $k<j\le n$. Then we have for this ν

$$\left|\sum_{j=k}^{n}I_\nu(\phi_{i,\kappa_s}(\omega^{j-1}))(\omega_j-p_j)\right|\ge(0.25-\kappa_s)(n-k), \quad (31)$$

where $p_j=\phi_i(\omega^{j-1})$, $i=p(p(n))$ and $s=p'(p(n))$. Since the ratio k/n and the number κ_s become arbitrary small for sufficiently large n, we have (18) for some selection rule δ, where δ satisfies $\delta(\omega^{j-1})=I_\nu(\phi_{i,\kappa_s}(\omega^{j-1}))$ for $k<j\le n$.

Sketch of the proof of Theorem 3. To prove (17), we define in (21) $\phi_i(x)=E_x$ - the mathematical expectation of the forecast generated by P_x.

Let κ_s be as in the proof of Theorem 4 and $E_{\omega^{j-1},s}=\phi_{i,\kappa_s}(\omega^{j-1})$.

Then in the proof of Lemma 6, for $\nu = 0$ or for $\nu = 1$, the inequality (31), where p_j is replaced by $E_{\omega^{j-1}}$, holds for infinitely many n. By the strong law of large numbers we obtain that, for $\nu = 0$ and for $\nu = 1$, with Pr-probability one

$$\frac{1}{n}\sum_{j=1}^{n} I_\nu(E_{\omega^{j-1},s})(p_j - E_{\omega^{j-1}}) \to 0$$

as $n \to \infty$, where i and s are defined as in the proof of Lemma 6. From this (17) can be easily obtained.

A Proof of Proposition 1: the universal randomized rounding algorithm

In this section we present a simplified version of Kakade and Foster's [7] randomized rounding algorithm. Assume $\omega = \omega_1\omega_2\dots$ is an infinite binary sequence given to a forecaster online. We define a randomized forecasting system $P_{\omega^{n-1}}$ such that the corresponding overall probability Pr of the event (2) tends to 1 as $n \to \infty$.

Let the rational points $v_i = i\Delta$, $i = 0, 1, \dots, K$, be divide the unit interval $[0, 1]$ on equal parts of length $\Delta = 1/K$. Denote $V = \{v_0, \dots, v_K\}$. Any real number $p \in [0, 1]$ is represented as a linear combination of two neighboring points of the corresponding partition

$$p = \sum_{v \in V} w_v(p)v = w_{v_{i-1}}(p)v_{i-1} + w_{v_i}(p)v_i,$$

where $p \in [v_{i-1}, v_i]$, $i = \lfloor p/\Delta + 1\rfloor$, and $w_{v_{i-1}}(p) = 1 - (p - v_{i-1})/\Delta$, $w_{v_i}(p) = 1 - (v_i - p)/\Delta$. Put $w_v(p) = 0$ for all other $v \in V$.

In that follows, some deterministic forecast p will be rounded to v_{i-1} with probability $w_{v_{i-1}}(p)$ or to v_i with probability $w_{v_i}(p)$. Now we define the deterministic algorithm computing this p. Assume the forecasts $p_1, \dots, p_{n-1}$ be already defined (put $p_1 = 0$). Define a real number p_n. For any $v \in V$, we consider an auxiliary function

$$\mu_{n-1}(v) = \sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i).$$

The weighted sum of these functions can be transformed as follows

$$\sum_{v\in V} w_v(p)\mu_{n-1}(v) = \sum_{v\in V} w_v(p)\sum_{i=1}^{n-1} w_v(p_i)(\omega_i - p_i) =$$
$$\sum_{i=1}^{n-1}(\sum_{v\in V} w_v(p)w_v(p_i))(\omega_i - p_i) =$$
$$\sum_{i=1}^{n-1}(\bar{w}(p)\cdot\bar{w}(p_i))(\omega_i - p_i) = \sum_{i=1}^{n-1} K(p,p_i)(\omega_i - p_i),$$

where $\bar{w}(p) = (0,\ldots,w_{v_{i-1}}(p), w_{v_i}(p),\ldots,0)$ is the vector of probabilities of rounding, $p\in[v_{i-1},v_i]$, and $K(p,p_i) = (\bar{w}(p)\cdot\bar{w}(p_i))$ is the dot product (kernel function). By definition $K(p,p_i)$ is a continuous function (it is piecewise linear).

We have

$$(\mu_n(v))^2 = (\mu_{n-1}(v))^2 + 2w_v(p_n)\mu_{n-1}(v)(\omega_n - p_n) + \\ +(w_v(p_n))^2(\omega_n - p_n)^2. \tag{A.1}$$

Summing (A.1) by v, we obtain

$$\sum_{v\in V}(\mu_n(v))^2 = \sum_{v\in V}(\mu_{n-1}(v))^2 + 2(\omega_n - p_n)\sum_{v\in V} w_v(p_n)\mu_{n-1}(v) + \\ +\sum_{v\in V}(w_v(p_n))^2(\omega_n - p_n)^2. \tag{A.2}$$

The second term of the sum (A.2) can be made less or equal to 0 for some p_n. Indeed, let p_n be equal to a root p of the equation

$$\sum_{v\in V} w_v(p)\mu_{n-1}(v) = \sum_{i=1}^{n-1} K(p,p_i)(\omega_i - p_i) = 0 \tag{A.3}$$

if a solution of (A.3) exists. Otherwise if the left part of the equation (A.3) (which is a continuous by p function) is positive for all p when put $p_n = 1$; put $p_n = 0$ if it is negative. By definition p_n is the deterministic forecast.

The third term of (A.2) is bounded by 1. Indeed,

$$\sum_{v\in V}(w_v(p_n))^2(\omega_n - p_n)^2 \le \sum_{v\in V} w_v(p_n) = 1.$$

Then by (A.2) for the forecasts p_i, $i = 1,\ldots,n$, computed above we have

$$\sum_{v\in V}(\mu_n(v))^2 \le \sum_{i=1}^{n}\sum_{v\in V}(w_v(p_i))^2(\omega_i - p_i)^2 \le n. \tag{A.4}$$

Since the sum of squares is less or equal than n, each member of this sum is less or equal than n. Hence, for any $v \in V$ we have

$$\frac{1}{n}\sum_{i=1}^{n} w_v(p_i)(\omega_i - p_i) \le \frac{1}{\sqrt{n}}. \tag{A.5}$$

Define $P_{\omega^{n-1}}(v) = w_v(p_i)$ for all $v \in V$ (by definition for any p, only two nearby values of $w_v(p)$ are nonzero).

Assume that $I(p)$ is the characteristic function of some subinterval of $[0,1]$. For each i, the mean value of the random variable $I(\tilde{p}_i)(\omega_i - \tilde{p}_i)$, where $\tilde{p}_i$ is distributed according to $P_{\omega^{n-1}}$, is equal to

$$E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) = \sum_{v\in V} w_v(p_i)I(v)(\omega_i - v).$$

By the martingale strong law of large numbers with Pr-probability 1

$$\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) - \frac{1}{n}\sum_{i=1}^{n} E(I(\tilde{p}_i)(\omega_i - \tilde{p}_i)) \to 0$$

as $n \to \infty$, where the overall probability Pr is defined by $P_{\omega^{i-1}}$, $i = 1, 2, \ldots$.

By definition of p_i and $w_v(p)$

$$\left|\sum_{v\in V} w_v(p_i)I(v)(\omega_i - v) - \sum_{v\in V} w_v(p_i)I(v)(\omega_i - p_i)\right| \le \Delta.$$

Since (A.5) holds for each $v \in V$ and the set V is finite,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{v\in V} w_v(p_i)I(v)(\omega_i - p_i) \to 0$$

as $n \to \infty$. Therefore, the Pr-probability of the event

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i)\right| \le \Delta$$

tends to 1 as $n \to \infty$.

References

[1] A.P.Dawid, The well-calibrated Bayesian [with discussion], J. Am. Statist. Assoc. 77 (1982) 605-613.

[2] A.P. Dawid, Calibration-based empirical probability [with discussion], Ann. Statist. 13 (1985) 1251-1285.

[3] A.P. Dawid, The impossibility of inductive inference, J. Am. Statist. Assoc. 80 (1985) 340-341.

[4] D.P. Foster, R. Vohra, Asymptotic calibration, Biometrika 85 (1998) 379-390.

[5] M. Hutter, New Error Bounds for Solomonoff Prediction, Journal of Computer and System Sciences 62 (4) (2001) 653-667.

[6] A. Chernov, M. Hutter, and J. Schmidhuber, Algorithmic complexity bounds on future prediction errors, Information and Computation, 205 (2) 2007, 242-261.

[7] S.M. Kakade, D.P. Foster, Deterministic calibration and Nash equilibrium, LNCS 3120 (John Shawe Taylor and Yoram Singer. ed) (2004) 33-48.

[8] Lehrer, E., Any Inspection Rule is Manipulable, Econometrica, 69-5, (2001) 1333-1347.

[9] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and Its Applications, Springer-Verlag. New York (1997).

[10] N. Merchav, M. Feder, Universal prediction, IEEE Trans. Inform. Theory, 44 (1998) 2124-2147.

[11] D. Oakes, Self-calibrating priors do not exist [with discussion], J. Am. Statist. Assoc. 80 (1985) 339-342.

[12] H. Rogers, Theory of recursive functions and effective computability, McGraw Hill. New York (1967).

[13] Sandroni, A., Smorodinsky R., and Vohra, R., Calibration with Many Checking Rules, Mathematics of Operations Research, 28-1, (2003) 141-153.

[14] Shafer, G., Vovk, V., Probability and Finance. It's Only a Game! New York: Wiley, 2001.

[15] Shiryaev, A.N.. Probability Berlin:Springer 1980.

[16] R.J. Solomonoff, A formal theory of inductive inference I, II, Inform. Control 7, (1964) 1-22, 224-254.

[17] R.J. Solomonoff, Complexity-based induction systems: Comparisons and convergence theorems, IEEE Trans. Inform. Theory, IT-24, (1978) 422-432.

[18] Vladimir Vovk, Glenn Shafer, Good randomized sequential probability forecasting is always possible, J. Royal Stat. Soc. B, 67 (2005) 747-763.

[19] Vladimir Vovk, Akimichi Takemura, Glenn Shafer, Defensive Forecasting, Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (2005) 365-372 (http://arxiv.org/abs/cs/0505083).

[20] V.V. V'yugin, Non-stochastic infinite and finite sequences, Theor. Comp. Science 207 (1998) 363-382.

[21] Vladimir V. V'yugin, On Calibration Error of Randomizing Forecasting Algorithms, ALT 2007 (M.Hutter, R.V.Servedio, and E.Takimoto (Eds.) LNAI 4754 380-394 Springer Verlag, Berlin - Heildeberg 2007.

[22] V.A. Uspensky, A.L. Semenov, A.Kh. Shen, Can an individual sequence of zeros and ones be random, Russian Math. Surveys 45(1) (1990) 121-189.

[23] A.K. Zvonkin, L.A. Levin, The complexity of finite objects and the algorithmic concepts of information and randomness, Russ. Math. Surv. 25 (1970) 83-124.