Calibration with Parameterized Selection Rules

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

Abstract

We generalize Dawid's notion of calibration for more general selection rules. Also, we extend Kakade and Foster's algorithm for arbitrary real valued outcomes and for this modified notion of calibration. Upper bounds for rate of convergence of calibrated forecasts are presented.

1 Introduction

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 an subjective estimate to future states.

A minimal requirement for testing any prediction algorithm is that it should be calibrated (see Dawid [2]). Dawid gave an informal explanation of calibration for binary outcomes 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.

A more precise definition is as follows. 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. We call such a function an indicator function. An infinite sequence of forecasts $p_1, p_2, \dots$ is calibrated for an infinite binary sequence of outcomes $\omega_1\omega_2\dots$ if for characteristic function $I(p)$ of any subinterval of $[0,1]$ the calibration error tends to zero, i.e.,

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

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.

If the weather acts adversatively, then Oakes [6] and Dawid [3] show that any deterministic forecasting algorithm will not always be calibrated.

Foster and Vohra [4] show that calibration is almost surely guaranteed with a randomizing forecasting rule, i.e., where the forecasts p_i 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 [5] presented "an almost deterministic" randomized rounding universal forecasting algorithm. For any sequence of outcomes and for any precision

*This research was partially supported by Russian foundation for fundamental research: 09-07-00180-a

of rounding $\Delta > 0$, an observer can simply randomly round the deterministic forecast p_i up to Δ in order to calibrate for this sequence with probability one :

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \le \Delta, \tag{1}$$

where $\tilde{p}_i$ is a random forecast. This algorithm can be easily extended such that the calibration error tends to zero as $n \to \infty$.

The goal of this paper is to extend Kakade and Foster's algorithm to arbitrary real valued outcomes $\omega_i \in [0,1]$ and to a more general notion of calibration with changing selection rules. A closely related approach for forecast continuous supermartingales is presented in Vovk [7]. We present also convergence bounds depending on the number of parameters.

2 Main result

Let $y_1, y_2, \dots$ be an infinite sequence of real numbers. An infinite sequence of random variables $\tilde{y}_1, \tilde{y}_2, \dots$ is called *a randomization* of $y_1, y_2, \dots$ if $E_n(\tilde{y}_n) = y_n$ for all n, where E_n is the symbol of mathematical expectation.

Computing forecasts, we can use a side information, or signals, $\bar{x}_1, \bar{x}_2, \dots$ given online: for any n, a k-dimensional vector $\bar{x}_n \in [0,1]^k$ is given to a forecaster before he announces his forecast $\tilde{p}_n$. We consider selection rules of a general type – indicator functions $I(p, \bar{x})$, where $p \in [0,1]$ and $\bar{x} \in [0,1]^k$.

An example of such indicator function useful for financial applications is

$$I(p_i, \omega_{i-1}) = \begin{cases} 1, & \text{if } p_i > \omega_{i-1} + \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Here $k = 1$, $\epsilon > 0$, and $x_i = \omega_{i-1}$ is the past outcome.

Theorem 1 *Given k an algorithm f for computing forecasts and a method of randomization can be constructed such that for any sequence of real numbers $\omega_1, \omega_2, \dots$ and for any sequence of k-dimensional signals $\bar{x}_1, \bar{x}_2, \dots$ the event*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i) = 0, \tag{2}$$

has Pr-probability 1, where Pr is a probability distribution generated by a sequence of tuples $(\tilde{p}_i, \tilde{x}_i)$ of random variables, $i = 1, 2, \dots$, and I is an arbitrary indicator function. Here $\tilde{p}_i$ is the randomization of a forecast p_i computed by the forecasting algorithm f and $\tilde{x}_i$ is obtained by independent randomization of each coordinate $x_{i,j}$ of the vector $\bar{x}_i$, $j = 1, \dots k$. Also $\mathrm{Var}_n(\tilde{p}_n) \to 0$ and $\mathrm{Var}_n(\tilde{x}_{i,j}) \to 0$ as to $n \to \infty$ for all i and j. [1]

Proof. We modify a weak calibration algorithm of Kakade and Foster [5] using also ideas from Vovk [7]. At first, we construct an Δ-calibrated forecasting algorithm, and after that we apply some double trick argument for it. We prove that given k an algorithm for computing forecasts and a method of randomization can be constructed such that for any sequence of real numbers $\omega, \omega_2, \dots$ and for any sequence of k-dimensional signals $\bar{x}_1, \bar{x}_2, \dots$ the event

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i) \le \Delta$$

[1] $\mathrm{Var}_n(\tilde{p}_n) = E_n(\tilde{p}_n - p_n)^2$.

has Pr-probability 1, where Pr and I is an arbitrary indicator function. Also $\mathrm{Var}_n(\tilde{p}_n) \le \Delta$ and $\mathrm{Var}_n(\tilde{x}_{i,j}) \le \Delta$ for all n, for all i and j.

At first we define a deterministic forecast and after that we randomize it.

Divide the interval $[0,1]$ on subintervals of length $\Delta = 1/K$ with rational endpoints $v_i = i\Delta$, where $i = 0, 1, \dots, K$. Let V denotes the set of these points.

Any number $p \in [0,1]$ can be represented as a linear combination of two neighboring endpoints of V defining subinterval containing p : $p = \sum\limits_{v\in V} w_v(p)v = w_{v_{i-1}}(p)v_{i-1} + w_{v_i}(p)v_i$, where $p \in [v_{i-1}, v_i]$, $i = \lfloor p^1/\Delta + 1 \rfloor$, $w_{v_{i-1}}(p) = 1 - (p - v_{i-1})/\Delta$, and $w_{v_i}(p) = 1 - (v_i - p)/\Delta$. Define $w_v(p) = 0$ for all other $v \in V$.

In that follows we round some deterministic forecast p_n to v_{i-1} with probability $w_{v_{i-1}}(p_n)$ and to v_i with probability $w_{v_i}(p_n)$. We also round the each coordinate $x_{n,s}$, $s = 1, \dots k$, of a signal $\bar{x}_n$ to v_{j_s-1} with probability $w_{v_{j_s-1}}(x_{n,s})$ and to v_{j_s} with probability $w_{v_{j_s}}(x_{n,s})$, where $x_{n,s} \in [v_{j_s-1}, v_{j_s}]$.

Let also $W_v(Q_n) = w_{v^1}(p_n)w_{v^2}(\bar{x}_n)$, where $v = (v^1, v^2)$, $v^1 \in V$, $v^2 = (v_1^2, \dots v_k^2) \in V^k$, $w_{v^2}(\bar{x}_n) = \prod_{s=1}^k w_{v_s^2}(x_{n,s})$, and $Q_n = (p_n, \bar{x}_n)$. For any Q_n, $W_v(Q_n)$ is a probability distribution in V^{k+1} : $\sum\limits_{v\in V^{k+1}} W_v(Q_n) = 1$.

In that follows we define a deterministic forecast p_n. Let the forecasts $p_1, \dots, p_{n-1}$ already defined (put $p_1 = 1/2$). Let us define for $v = (v^1, v^2)$ and $Q_i = (p_i, \bar{x}_i)$

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

We have

$$\begin{aligned} (\mu_n(v))^2 &= (\mu_{n-1}(v))^2 + \\ +2W_v(Q_n)\mu_{n-1}(v)(\omega_n - p_n) &+ (W_v(Q_n))^2(\omega_n - p_n^1)^2. \end{aligned} \tag{3}$$

Summing (3) by $v \in V^{k+1}$, we obtain:

$$\begin{aligned} \sum_{v\in V^{k+1}} (\mu_n(v))^2 &= \sum_{v\in V^{k+1}} (\mu_{n-1}(v))^2 + \\ +2(\omega_n - p_n) \sum_{v\in V^{k+1}} W_v(Q_n)\mu_{n-1}(v) &+ \sum_{v\in V^{k+1}} (W_v(Q_n))^2(\omega_n - p_n)^2. \end{aligned} \tag{4}$$

Change the order of summation:

$$\begin{aligned} \sum_{v\in V^{k+1}} W_v(Q_n)\mu_{n-1}(v) &= \sum_{v\in V^{k+1}} W_v(Q_n) \sum_{i=1}^{n-1} W_v(Q_i)(\omega_i - p_i) = \\ &= \sum_{i=1}^{n-1} \Big(\sum_{v\in V^{k+1}} W_v(Q_n)W_v(Q_i)\Big)(\omega_i - p_i) = \\ = \sum_{i=1}^{n-1} (\bar{W}(Q_n) \cdot \bar{W}(Q_i))(\omega_i - p_i) &= \sum_{i=1}^{n-1} K(Q_n, Q_i)(\omega_i - p_i), \end{aligned}$$

where $\bar{W}(Q_n) = (W_v(Q_n) : v \in V^{k+1})$, $\bar{W}(Q_n) = (W_v(Q_n) : v \in V^{k+1})$ be vectors of probabilities of rounding. The dot product of corresponding vectors defines the kernel

$$K(Q_n, Q_i) = K(p_n, \bar{x}_n, p_i, \bar{x}_i) = (\bar{W}(Q_n) \cdot \bar{W}(Q_i)). \tag{5}$$

Let p_n be equal to the root of the equation

$$S_n(p_n) = \sum_{v\in V} W_v(p_n, \bar{x}_n)\mu_{n-1}(v) = \sum_{i=1}^{n-1} K(p_n, \bar{x}_n, p_i, \bar{x}_i)(\omega_i - p_i) = 0, \tag{6}$$

if a solution exists. Otherwise, if the left hand-side of the equation (6) (which is a continuous by p_n function) strictly positive for all p_n define $p_n = 1$, define $p_n = 0$ if it is strictly negative. Announce p_n as a deterministic forecast.

The third term of (4) is upper bounded by 1. Indeed, since $|\omega_i - p_i| \le 1$ for all i,

$$\sum_{v \in V^{k+1}} (W_v(Q_n))^2(\omega_i - p_n)^2 \le \sum_{v \in V^{k+1}} W_v(Q_n) = 1.$$

Then by (4), $\sum\limits_{v \in V^{k+1}} (\mu_n(v))^2 \le n$. Recall that for any $v \in V^{k+1}$

$$\mu_n(v) = \sum_{i=1}^{n} W_v(Q_i)(\omega_i - p_i). \tag{7}$$

Insert the term $I(v)$ in the sum (7), where I is an arbitrary indicator function and $v \in [0,1]^{k+1}$, sum by $v \in V^{k+1}$, and exchange the order of summation. Using Cauchy–Schwartz inequality for vectors $(I(v) : v \in V^{k+1})$, $(\mu_n(v) : v \in V^{k+1})$ and Euclidian norm, we obtain

$$\left| \sum_{i=1}^{n} \sum_{v \in V^{k+1}} W_v(Q_i) I(v)(\omega_i - p_i) \right| =$$

$$= \left| \sum_{v \in V^{k+1}} I(v) \sum_{i=1}^{n} W_v(Q_i)(\omega_i - p_i) \right| \le$$

$$\le \sqrt{\sum_{v \in V^{k+1}} I(v)} \sqrt{\sum_{v \in V^{k+1}} (\mu_n(v))^2} \le \sqrt{|V^{k+1}| n} \tag{8}$$

for all n, where $|V^{k+1}| = 1/\Delta^{k+1}$ – is the cardinality of the partition.

Let $\tilde{p}_i$ be a random variable taking values $v \in V$ with probabilities $w_v(p_i)$ (only two of them are nonzero). Recall that $\tilde{x}_i$ is a random variable taking values $v \in V^k$ with probabilities $w_v(\bar{x}_i)$.

Let I be an indicator function of $k+1$ arguments. For any i the mathematical expectation of a random variable $I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i)$ is equal to

$$E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i)) = \sum_{v \in V^{k+1}} W_v(Q_i) I(v)(\omega_i - v^1), \tag{9}$$

where $v = (v^1, v^2)$.

By the strong law of large numbers, for some $\mu_n = o(n)$ (as $n \to \infty$), *Pr*-probability of the event

$$\left| \sum_{i=1}^{n} I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i) - \sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i)) \right| \le \mu_n \tag{10}$$

tends to 1 as $n \to \infty$. A form of of μ_n will be specified later.

By definition of deterministic forecast

$$\left| \sum_{v \in V^{k+1}} W_v(Q_i) I(v)(\omega_i - p_i) - \sum_{v \in V^{k+1}} W_v(Q_i) I(v)(\omega_i - v^1) \right| < \Delta$$

for all i, where $v = (v^1, v^2)$. Summing (9) by $i = 1, \dots, n$ and using the inequality (8), we obtain

$$\left| \sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i)) \right| =$$

$$= \left| \sum_{i=1}^{n} \sum_{v \in V^{k+1}} W_v(Q_i) I(v)(\omega_i - v^1) \right| < \Delta n + \sqrt{|V^{k+1}| n} \tag{11}$$

for all n, where $|V^{k+1}| = 1/\Delta^{k+1}$ is the cardinality of the partition.

By (11) and (10) we obtain that Pr-probability of the event

$$\left|\sum_{i=1}^{n} I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i)\right| \le \Delta n + \mu_n + \sqrt{n/\Delta^{k+1}} \tag{12}$$

tends to 1 as $n\to\infty$. In particular, Pr-probability of the event

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

is equal to 1.

To prove that (2) holds almost surely choose a monotonic sequence of rational numbers $\Delta_1 > \Delta_2 > \dots$ such that $\Delta_s \to 0$ as $s\to\infty$. We also define an increasing sequence of natural numbers $n_1 < n_2 < \dots$ For any s, we use on steps $n_s \le n < n_{s+1}$ the partition of $[0,1]$ on subintervals of length Δ_s.

We choose n_s such that $n_s \ge \left(\frac{k+2}{2}\right)^2 \Delta_s^{-(k+3)}$ for all s. [2] Put $n_0 = 0$ and $\Delta_0 = 1$. Also, define the numbers $n_1, n_2, \dots$ such that the inequality

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i))\right| \le 4(s+1)\Delta_s n \tag{13}$$

holds for all $n_s \le n \le n_{s+1}$ and for all s.

We define this sequence by mathematical induction on s. Suppose that n_s $(s\ge 1)$ is defined such that the inequality

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i))\right| \le 4s\Delta_{s-1} n \tag{14}$$

holds for all $n_{s-1} \le n \le n_s$, and the inequality

$$\left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i))\right| \le 4s\Delta_s n_s \tag{15}$$

also holds. Let us define n_{s+1}. Consider all forecasts $\tilde{p}_i$ defined by the algorithm given above for discretization $\Delta = \Delta_{s+1}$. We do not use first n_s of these forecasts (more correctly we will use them only in bounds (16) and (17); denote these forecasts $\mathbf{\hat{p}_1}, \dots, \mathbf{\hat{p}_{n_s}}$). We add the forecasts $\tilde{p}_i$ for $i > n_s$ to the forecasts defined before this step of induction (for n_s). Let n_{s+1} be such that the inequality

$$\begin{aligned}\left|\sum_{i=1}^{n_{s+1}} E(I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i))\right| &\le \left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i))\right| + \\ + \left|\sum_{i=n_s+1}^{n_{s+1}} E(I(\tilde{p}_i,\tilde{x}_i)(\omega_i-\tilde{p}_i)) + \sum_{i=1}^{n_s} E(I(\mathbf{\hat{p}_i},\tilde{x}_i)(\omega_i-\mathbf{\hat{p}_i}))\right| &+ \\ + \left|\sum_{i=1}^{n_s} E(I(\mathbf{\hat{p}_i},\tilde{x}_i)(\omega_i-\mathbf{\hat{p}_i}))\right| &\le 4(s+1)\Delta_{s+1} n_{s+1}\end{aligned} \tag{16}$$

holds. Here the first sum of the right-hand side of the inequality (16) is bounded by $4s\Delta_s n_s$ – by the induction hypothesis (15). The second and third sums are bounded

[2] This is the minimum point of (11).

by $2\Delta_{s+1}n_{s+1}$ and by $2\Delta_{s+1}n_s$, respectively. This follows from (11) and by choice of n_s. The induction hypothesis (15) is valid for

$$n_{s+1} \ge \frac{2s\Delta_s + \Delta_{s+1}}{\Delta_{s+1}(2s+1)} n_s.$$

Analogously,

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i))\right| \le \left|\sum_{i=1}^{n_s} E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i))\right| +$$
$$+\left|\sum_{i=n_s+1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i)) + \sum_{i=1}^{n_s} E(I(\hat{\mathbf{p}}_{\mathbf{i}}, \tilde{x}_i)(\omega_i - \hat{\mathbf{p}}_{\mathbf{i}}))\right| +$$
$$+\left|\sum_{i=1}^{n_s} E(I(\hat{p}_i, \tilde{x}_i)(\omega_i - \hat{\mathbf{p}}_{\mathbf{i}}))\right| \le 4(s+1)\Delta_s n \quad (17)$$

for $n_s < n \le n_{s+1}$. Here the first sum of the right-hand inequality (16) is also bounded by $4s\Delta_s n_s \le 4s\Delta_s n$ – by the induction hypothesis (15). The second and the third sums are bounded by $2\Delta_{s+1}n \le 2\Delta_s n$ and by $2\Delta_{s+1}n_s \le 2\Delta_s n$, respectively. This follows from (11) and from choice of Δ_s. The induction hypothesis (14) is valid.

By (13) for any s

$$\left|\sum_{i=1}^{n} E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i))\right| \le 4(s+1)\Delta_s n \quad (18)$$

for all $n \ge n_s$ if Δ_s satisfies the condition $\Delta_{s+1} \le \Delta_s(1 - \frac{1}{s+2})$ for all s.

By the law of large numbers (22), the relation (10) can be specified:

$$Pr\left\{\sup_{n \ge n_s}\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right| > \Delta_s\right\} \le (\Delta_s)^{-2}e^{-2n_s\Delta_s^2} \quad (19)$$

for all s, where $V_i = I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i) - E(I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i))$ is a sequence of martingale–differences.

Combining (18) with (19), we obtain

$$Pr\left\{\sup_{n \ge n_s}\left|\frac{1}{n}\sum_{i=1}^{n} I(\tilde{p}_i, \tilde{x}_i)(\omega_i - \tilde{p}_i)\right| \ge (4s+5)\Delta_s\right\} \le (\Delta_s)^{-2}e^{-2n_s\Delta_s^2} \quad (20)$$

for all s. The series $\sum_{s=1}^{\infty}(\Delta_s)^{-2}e^{-2n_s\Delta_s^2}$ is convergent if n_s satisfies

$$n_s \ge \frac{\ln s + 2\ln\ln s - 2\ln(\Delta_s)}{2\Delta_s^2}$$

for all s. Let also $\Delta_s = o(1/s)$ as $s \to \infty$. Then Borel–Cantelli Lemma implies convergence of (2) almost surely.

It is easy to verify that the sequences n_s and Δ_s satisfying all the conditions above exist.

Let us specify details of rounding. The expression $\Delta n + \sqrt{n/\Delta^{k+1}}$ from (12) takes its minimal value for $\Delta = (\frac{k+1}{2})^{\frac{2}{k+3}} n^{-\frac{1}{k+3}}$. In this case, the right-hand side of the inequality (11) is equal to $\Delta n + \sqrt{n/\Delta^{k+1}} = 2\Delta n = 2(\frac{k+1}{2})^{\frac{2}{k+3}} n^{1-\frac{1}{k+3}}$.

For example, for $k = 1$, we use at any step n the rounding $\Delta_s = n_s^{-1/4}$, where s is such that $n_s < n \le n_{s+1}$.

We write $A \sim B$ if positive constants c_1 and c_2 exist such that $c_1 B \leq A \leq c_2 B$ for all values of parameters from the expressions A and B.

Define $n_s = s^M$ and $\Delta_s = s^{-M/4}$, where M is a positive integer number. Then $s \sim n_s^{1/M}$ (the constants c_1 and c_2 depend on M).

Easy to verify that all requirements for n_s and Δ_s given in Section 2 are valid.

By (20) we can define $\mu_n = (4s+5)\Delta_s n$, where s is such that $n_s < n \leq n_{s+1}$. For $n_s < n \leq n_{s+1}$ it holds $n \sim n_s$, hence, $\mu_n \sim n^{3/4+1/M}$.

A Large deviation inequality for martingales

A sequence $V_1, V_2, \dots$ is called martingale-difference with respect to a sequence of random variables $X_1, X_2, \dots$ if for any $i > 1$ the random variable V_i is a function of $X_1, \dots, X_i$ and $E(V_{i+1}|X_1, \dots, X_i) = 0$ almost surely. The following inequalities are corollaries of Hoffding-Azuma inequality [1]:

Let $V_1, V_2, \dots$ be a martingale–difference with respect to $X_1, X_2, \dots$, and $V_i \in [A_i, A_i + 1]$ for some random variable A_i measurable with respect to $X_1, \dots, X_i$. Let $S_n = \sum\limits_{i=1}^{n} V_i$. Then for any $t > 0$

$$P\left\{ \left| \frac{S_n}{n} \right| > t \right\} \leq 2e^{-2nt^2} \tag{21}$$

for all n. A strong law of large numbers is also holds: for any t

$$P\left\{ \sup_{k \geq n} \left| \frac{S_k}{k} \right| > t \right\} \leq t^{-2} e^{-2nt^2} \tag{22}$$

for all n. Since the series of the exponents from the right-hand side of the inequality (21) convergent, by Borel–Cantelli Lemma we obtain the martingale strong law of large numbers

$$P\left\{ \lim_{n \to \infty} \frac{S_n}{n} = 0 \right\} = 1.$$

References

[1] N.Cesa-Bianchi, G.Lugosi. Prediction, Learning and Games. Cambridge University Press, 2006.

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

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

[4] D.P.Foster, R. Vohra. Asymptotic calibratio. Biometrika. 85, 379–390, 1998.

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

[6] D.Oakes. Self-Calibrating Priors Do not Exist [with discussion]. J. Am. Statist. Assoc. V.80. P.339–342, 1985.

[7] V.Vovk. Defensive Forecasting for Optimal Prediction with Expert Advice. arXiv:0708.1503v1. 2007.