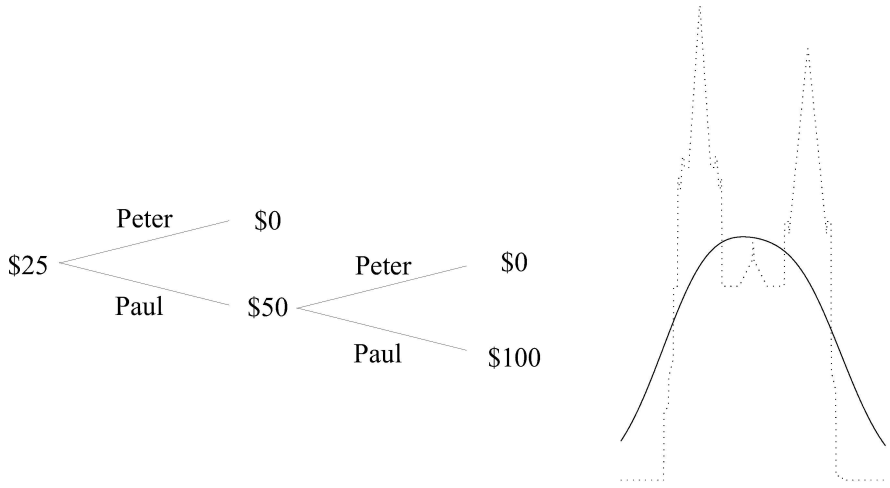
A law of large numbers for predicting several steps ahead

Vladimir Vovk



The Game-Theoretic Probability and Finance Project

Working Paper #67

First posted August 24, 2025. Last revised August 27, 2025.

Project web site:
http://www.probabilityandfinance.com

Abstract

This note proves a law of large numbers for predicting several steps ahead, which, in the case of uniformly bounded random variables, generalizes the standard law of large numbers for martingales; the standard law of large numbers corresponds to predicting one step ahead. Its main result shows that the law of large numbers holds for predicting N uniformly bounded random variables $o(N)$ steps ahead, but it is much more precise and in some respects optimal. This law of large numbers is applied to a problem of decision making with a bounded loss function limiting the impact of each decision to $o(N)$ steps.

This note has also been published as an arXiv report.

Contents

1 Introduction

The usual martingale statements of limit theorem of probability theory involve one-step-ahead averages $\mathbb{E}(Y_n \mid \mathcal{F}_{n-1})$, where Y_n is an adapted sequence for a filtration $(\mathcal{F}_n)$. A typical law of large numbers says that, under some conditions (such as Y_n being uniformly bounded),

$$\frac{1}{N}\sum_{n=1}^{N}(Y_n - \mathbb{E}(Y_n \mid \mathcal{F}_{n-1})) \approx 0 \tag{1}$$

with high probability, provided N is large enough. This property may be expressed by saying that the one-step-ahead forecasts $\mathbb{E}(Y_n \mid \mathcal{F}_{n-1})$ for Y_n are asymptotically unbiased. It is easy to see that (1) continues to hold for predicting K steps ahead when K is a small constant:

$$\frac{1}{N}\sum_{n=1}^{N}(Y_n - \mathbb{E}(Y_n \mid \mathcal{F}_{n-K})) \approx 0. \tag{2}$$

The question asked in this note is how large K can be in order to have (2) with high probability for uniformly bounded Y_n. A crude answer is that, for uniformly bounded Y_n, it can be of the order $o(N)$ but no more in general.

Section 2 states main positive result of this note, while Sections 3 and 4 explore its optimality. To demonstrate the usefulness of our law of large numbers, Section 5 applies it to a simple problem of decision making. Section 6 concludes and lists some directions of further research.

2 A law of large numbers in the form of inequality

We consider a two-sided *filtration* $(\mathcal{F}_n)$, $n \in \mathbb{Z} := \{\dots, -1, 0, 1, \dots\}$, on a probability space $(\Omega, \mathcal{F}, P)$; it is an increasing sequence of sub-σ-algebras of $\mathcal{F}$, i.e., $\mathcal{F}_n \subseteq \mathcal{F}_{n'}$ whenever $n \le n'$. As usual, a sequence (Y_n) of random variables in $(\Omega, \mathcal{F}, P)$ is *adapted* if Y_n is $\mathcal{F}_n$-measurable for all n. For simplicity we consider uniformly bounded Y_n and take, without loss of generality, 1 to be an upper bound for $|Y_n|$.

Theorem 2.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space equipped with a filtration $(\mathcal{F}_n)$, $n \in \mathbb{Z}$. Fix $K \in \{1, 2, \dots\}$. Let $Y_1, \dots, Y_N$ be an adapted sequence of random variables in $(\Omega, \mathcal{F}, P)$ bounded by 1 in absolute value, $|Y_n| \le 1$ for $n = 1, \dots, N$. Then we have, for any $\epsilon \in (0, 0.7)$,*

$$P\left(\left|\sum_{n=1}^{N}(Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}))\right| \ge 4\sqrt{K(N+K)\ln\frac{1}{\epsilon}}\right) < \epsilon. \tag{3}$$

We will sometimes refer to K as our *prediction horizon*. Our interpretation of (3) is that K-steps-ahead forecasts $\mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})$ for Y_n are asymptotically

unbiased when $K \ll N$. Our proof will use the primitive idea of decomposing forecasting K steps ahead into K processes of forecasting one step ahead, each of the K processes paying attention only to every Kth observation Y_n. Interestingly, this will give nearly optimal results, as we will see in Sections 3 and 4. To merge performance guarantees for the K processes we will need one result from robust risk aggregation, namely, [1, Theorem 4.2].

Proof of Theorem 2.1. It suffices to prove that

$$P\left(\left|\sum_{n=1}^{N}(Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}))\right| \geq 4\sqrt{KN\ln\frac{1}{\epsilon}}\right) < \epsilon \tag{4}$$

under the assumption that N is divisible by K. Indeed, if N is not, we can replace it by $N' := \lceil N/K \rceil K$ and apply (4) to N' in place of N.

We will need the following special case of Theorem 4.2 in [1]. Suppose nonnegative random variables X_k, $k = 1, \dots, K$, satisfy

$$\mathbb{P}(X_k \geq x) = \exp(-ax^2) \tag{5}$$

for all $x \geq 0$, where a is a positive constant. The value E of the optimization problem

$$\mathbb{P}(X_1 + \cdots + X_K \geq C) \to \max$$

(the max, or at least sup, being over all joint distributions for $(X_1, \dots, X_K)$ with the given marginals) does not exceed

$$\inf_{t<C/K} \frac{K\int_t^{C-(K-1)t}\exp(-ax^2)\,\mathrm{d}x}{C-Kt} \leq \inf_{t<C/K} \frac{K\int_t^{\infty}\exp(-ax^2)\,\mathrm{d}x}{C-Kt}. \tag{6}$$

We can extend the statement in the previous paragraph to a wider class of random variables X_k, $k = 1, \dots, K$. Namely, it suffices to assume that they satisfy

$$\mathbb{P}(X_k \geq x) \leq \exp(-ax^2) \tag{7}$$

for all $x \geq 0$, instead of (5) (if needed, we can increase such X_k, perhaps extending the underlying probability space, to make sure (5) holds). We will apply the statement to the random variables X_k given by

$$X_k := \sum_{n \in \{k, k+K, k+2K, \dots, k+(N/K-1)K\}} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}))\,.$$

By Hoeffding's inequality (see, e.g., [5, Section A.6.3]), for any $C > 0$ and any $k \in \{1, \dots, K\}$,

$$P\left(X_k \geq C\right) \leq \exp\left(-C^2/(2N/K)\right).$$

Therefore, (7) holds with

$$a := \frac{K}{2N}. \tag{8}$$

Let us set $t := \frac{C}{2K}$ in (6) (this is the middle of the range of t). This gives the upper bound

$$\frac{2K}{C} \int_{\frac{C}{2K}}^{\infty} \exp(-ax^2)\,\mathrm{d}x$$

for E, which can be rewritten (see below for an explanation) as

$$\frac{2K}{C} \frac{1}{\sqrt{2a}} \int_{\sqrt{2a}\frac{C}{2K}}^{\infty} \exp(-y^2/2)\,\mathrm{d}y = \frac{2K}{C} \frac{\sqrt{2\pi}}{\sqrt{2a}} \bar{\Phi}\left(\sqrt{2a}\frac{C}{2K}\right) \tag{9}$$

$$= \frac{2\sqrt{2\pi}\sqrt{KN}}{C} \bar{\Phi}\left(\frac{C}{2\sqrt{KN}}\right) < \frac{4KN}{C^2} \exp\left(-\frac{C^2}{8KN}\right). \tag{10}$$

The first expression in (9) is obtained by the substitution $y := \sqrt{2a}x$, the equality in (9) uses the notation $\bar{\Phi}$ for the survival function of the standard Gaussian distribution, the following equality (the one in (10)) is obtained by plugging in (8), and the final inequality in (10) follows from the usual upper bound for $\bar{\Phi}$ [2, Lemma VII.1.2].

To find a suitable solution to the inequality

$$\frac{4KN}{C^2} \exp\left(-\frac{C^2}{8KN}\right) < \frac{\epsilon}{2}, \tag{11}$$

we plug in $C = \sqrt{8KNx \ln \frac{1}{\epsilon}}$ (intuitively, x should not be so different from 1) obtaining, after simplification,

$$\epsilon^{x-1} < x \ln \frac{1}{\epsilon}.$$

Assuming $\epsilon < 0.7$, we can set $x := 2$, which gives (4). □

Remark 2.2. A typical statement of the law of large numbers replaces the sum in (3) by $\sum_{n=1}^{N} Y_n$ where Y_n satisfy $\mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}) = 0$. However, such a replacement would weaken Theorem 2.1. Indeed, while (3) immediately implies

$$(\forall n : \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) \Longrightarrow P\left(\left|\sum_{n=1}^{N} Y_n\right| \geq 4\sqrt{K(N+K)\ln\frac{1}{\epsilon}}\right) < \epsilon, \tag{12}$$

(12) does not imply (3), since centring a random variable (by subtracting its conditional expectation) can change its range $[-1, 1]$.

Remark 2.3. The proof of Theorem 2.1 also demonstrates the following one-sided counterpart of (12):

$$P\left(\sum_{n=1}^{N} Y_n \geq 4\sqrt{K(N+K)\ln\frac{1}{\epsilon}}\right) < \frac{\epsilon}{2}$$

provided $\mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}) \leq 0$ for all n a.s. Indeed, the inequality (11) shows that (3) continues to hold when the vertical bars are dropped in $|\ldots|$ and the ϵ on the right-hand side is replaced by $\frac{\epsilon}{2}$; we can then remove $\mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})$ as $\mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}) \leq 0$ a.s.

Remark 2.4. In the proof of Theorem 2.1 we did not make any attempt to optimize the coefficient 4 in (3). However, the same argument shows that 4 can be replaced by a number as close to $\sqrt{2}$ as we wish if we narrow down the permitted range of ϵ (leaving the lower end of the range at 0, of course).

3 An asymptotic statement and its optimality

Let us state Theorem 2.1 in a cruder way (traditional for stating the law of large numbers). According to (3),

$$\left|\sum_{n=1}^{N}(Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}))\right| = O_p\left(\sqrt{KN}\right) \tag{13}$$

(which assumes $K \le N$). The usual statement of the law of large numbers

$$\left|\frac{1}{N}\sum_{n=1}^{N}(Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K}))\right| = o_p(1)$$

follows when $K = o(N)$.

The following proposition is an inverse to (13).

Proposition 3.1. *Suppose the underlying probability space $(\Omega, \mathcal{F}, P)$ is atomless. For any N and any prediction horizon $K \le N$ there exists a sequence $Y_1, \dots, Y_N$ of random variables that are bounded by 1 in absolute values, $|Y_n| \le 1$ for $n = 1, \dots, N$, such that*

$$P\left(\sum_{n=1}^{N}(Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) \ge \sqrt{K(N-K)}\right) \ge 0.1, \tag{14}$$

where $(\mathcal{F}_n)$ is the filtration generated by (Y_n) (i.e., $\mathcal{F}_n := \sigma(Y_1, \dots, Y_n)$, meaning $\mathcal{F}_n := \{\emptyset, \Omega\}$ for $n < 1$).

The proof will shows that, when we replace the sum in (14) by its absolute value, as in (3), we can replace the 0.1 on the right-hand side of (14) by 0.2.

Proof of Proposition 3.1. Let us assume that N is divisible by K and prove

$$P\left(\sum_{n=1}^{N}(Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) \ge \sqrt{KN}\right) \ge 0.1 \tag{15}$$

(this will then imply (14) without this restriction by applying (15) to $N' := \lfloor N/K \rfloor K$ in place of N). Set $m := N/K$. Fix independent $\{-1, 1\}$-valued variables $X_1, \dots, X_m$ taking values ± 1 with equal probabilities (they exist by, e.g., [6, Lemma D.1 in the Online Supplement]), and set

$$Y_n := X_{\lceil n/K \rceil}, \quad n = 1, \dots, N.$$

Therefore, the N steps are split into m blocks of length K, and Y_n is constant within each block. By the central limit theorem applied to $X_1, \dots, X_m$, we have $Y_n = 1$ in at least $\sqrt{m}$ more blocks than $Y_n = -1$ with probability $\Phi(-1) \approx 0.159$ in the limit as $m \to \infty$, where Φ is the standard Gaussian distribution function. The smallest value of this probability for finite m is smaller, approximately 0.109, and it is attained at $m = 6$. If there is such an imbalance of at least $\sqrt{m}$,

$$\sum_{n=1}^{N} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) = \sum_{n=1}^{N} Y_n \geq K\sqrt{m} = K\sqrt{N/K} = \sqrt{KN},$$

which agrees with (15). □

4 Optimality in ϵ

Proposition 3.1 states the optimality of the dependence of the right-hand sides of (13) and of the inner inequality of (3) on K. Another natural question is whether the dependence of the right-hand side of the inner inequality of (3) on ϵ is optimal. A positive answer is provided by the following result, which is, however, somewhat more difficult to state and interpret than Proposition 3.1.

Proposition 4.1. *Suppose the underlying probability space is atomless. For any N, any prediction horizon K, and any $\epsilon > 0$ such that*

$$m := \frac{N}{K} \in 2\mathbb{Z}, \quad \sqrt{m \ln \frac{1}{15\epsilon}} \in 4\mathbb{Z}, \tag{16}$$

and

$$\frac{1}{2}\sqrt{KN \ln \frac{1}{15\epsilon}} \leq \frac{N}{4}, \tag{17}$$

there exists a sequence $Y_1, \dots, Y_N$ of random variables that are bounded by 1 in absolute values and satisfy

$$P\left(\sum_{n=1}^{N} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) \geq \frac{1}{2}\sqrt{KN \ln \frac{1}{15\epsilon}}\right) \geq \epsilon, \tag{18}$$

where $(\mathcal{F}_n)$ in (18) is the filtration generated by (Y_n).

If we ignore numerical constants such as 4 in (3) and $\frac{1}{2}$ in (18), the inequality (18) can be considered to be an inverse to (3). When (17) is violated, the inner inequality in (18) expresses an extreme bias of the predictions, namely, it implies

$$\frac{1}{N}\sum_{n=1}^{N} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) > \frac{1}{4};$$

the condition (17) thus means that our result falls short of dealing with such an extreme bias. The conditions (16) still leave us with a fairly dense net of permitted triples (N, K, ϵ).

Proof of Proposition 4.1. We will modify the proof of Proposition 3.1 by applying a lower bound for large deviations in the form of [4, Proposition 7.3.2]. Define $X_1, \dots, X_m$ and $Y_1, \dots, Y_N$ as before. Let Z be the number of times that $X_i = 1$, $i = 1, \dots, m$. Proposition 7.3.2 in [4] then says that

$$\mathbb{P}\left(Z \geq \frac{m}{2} + t\right) \geq \frac{1}{15} \exp\left(-16t^2/m\right) \tag{19}$$

provided m is even and t is an integer in the range $[0, m/8]$. Since

$$\sum_{n=1}^{N} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) = (2Z - m)K,$$

we can rewrite (19) as

$$\mathbb{P}\left(\sum_{n=1}^{N} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) \geq 2Kt\right) \geq \frac{1}{15} \exp\left(-16t^2/m\right). \tag{20}$$

If $\epsilon > 0$ is equal to the right-hand side of (20), we have

$$t = \frac{1}{4}\sqrt{m \ln \frac{1}{15\epsilon}},$$

and plugging thus expression for t into (20) gives (18).

The two conditions in [4, Proposition 7.3.2] are reflected in our statement: $t \leq m/8$ becomes (17), and m being even and t being an integer are required in (16). □

Remark 4.2. Kunsch and Rudolf [3, Lemma 3] slightly improve the constants in [4, Proposition 7.3.2], and using their result allows us to rewrite (18) in the form

$$\mathbb{P}\left(\sum_{n=1}^{N} (Y_n - \mathbb{E}_P(Y_n \mid \mathcal{F}_{n-K})) \geq 0.6\sqrt{KN \ln \frac{1}{4.3\epsilon}}\right) \geq \epsilon,$$

with the corresponding modifications of the conditions (16) and (17).

5 An application to decision making

In this section we will apply Theorem 2.1 to a problem of decision making. At each step n, $n = 1, 2, \dots$, a decision maker is required to make a decision $d \in \mathbf{D}$, chosen from a measurable space $(\mathbf{D}, \mathcal{D})$. The loss suffered by the decision maker as result of making the decision d at step n is measured by a measurable *loss function* $\lambda_n : \Omega \times \mathbf{D} \to [0, 1]$. The loss functions λ_n are assumed to be uniformly bounded, and without loss of generality, we take their range to be $[0, 1]$. Our notation for the random loss resulting from decision d taken at step n is $\lambda_n(d)$, so that $\lambda_n(d)(\omega) := \lambda_n(\omega, d)$.

We are interested in suitable strategies for the decision maker, referring to them as decision strategies. Formally, a *decision strategy* A is a sequence of adapted $\mathbf{D}$-valued random elements A_n; namely, each $A_n : \Omega \to \mathbf{D}$ is assumed to be $\mathcal{F}_n/\mathcal{D}$-measurable. Let us assume that there exists a *Bayesian strategy* B, i.e., a decision strategy satisfying

$$\mathbb{E}(\lambda_n(B_n) \mid \mathcal{F}_n) \le \mathbb{E}(\lambda_n(A_n) \mid \mathcal{F}_n) \quad \text{a.s.}$$

for any n and any other decision strategy A. When $\mathbf{D}$ is finite, the existence of a Bayesian strategy is automatic: we can define

$$B_n \in \arg\min_{d \in \mathbf{D}} \mathbb{E}(\lambda_n(d) \mid \mathcal{F}_n),$$

selecting the first decision d in a fixed ordering of $\mathbf{D}$ if the arg min contains more than one element. But in general, the existence of a Bayesian strategy has to be assumed. In interesting cases it usually exists and chooses an optimal decision at each step. The total loss of a decision strategy A over the first N steps is denoted by

$$\mathrm{Loss}_N(A) := \sum_{n=1}^{N} \lambda_n(A_n).$$

In order to show that the Bayesian strategy is better, or at least not much worse, than any alternative decision strategy, we have to assume that the loss resulting from the decision d made at step n does not depend substantially on remote future (if it does, our task is in general hopeless). Namely, we assume that $\lambda_n(d)$ becomes determined at step $n+K$ for some parameter $K \in \{1, 2, \dots\}$, which we will call the *impact horizon*, so that the effect of d disappears after step $n + K$; the role of the impact horizon is analogous to that of a prediction horizon. Formally, $\lambda_n : \Omega \times \mathbf{D} \to [0, 1]$ is assumed to be $\mathcal{F}_{n+K} \times \mathcal{D}$-measurable.

Corollary 5.1. *The Bayesian strategy B satisfies, for any $\epsilon \in (0, 0.7)$, any decision strategy A, and any N,*

$$P\left(\mathrm{Loss}_N(B) - \mathrm{Loss}_N(A) \ge 4\sqrt{K(N+K)\ln\frac{1}{\epsilon}}\right) < \epsilon. \tag{21}$$

According to (21), the Bayesian strategy's total loss over the first N steps (the total loss being determined only after $n+K$ steps) is, with high probability, almost as low as any other decision strategy's for a short impact horizon, $K \ll N$.

Proof of Corollary 5.1. Let us set $Y_{n+K} := \lambda_n(B_n) - \lambda_n(A_n)$, where we use the notation Y_{n+K} rather than Y_n to make this random sequence adapted. By the definition of a Bayesian strategy, we have $\mathbb{E}_P(Y_{n+K} \mid \mathcal{F}_n) \le 0$. Applying the one-sided law of large numbers as stated in Remark 2.3 but with $Y_{1+K}, \dots, Y_{n+K}$ in place of $Y_1, \dots, Y_n$, we obtain (21) (even with $\epsilon/2$ in place of ϵ on the right-hand side). □

6 Conclusion

This note proves a version of the law of large numbers for predicting several steps ahead and applies it to a problem of decision making with a limited impact horizon. These are some obvious directions of further research:

- get rid of the assumption that the random variables being averaged are uniformly bounded;
- find the optimal numerical constants in Theorem 2.1;
- establish tighter lower bounds corresponding to the law of large numbers;
- establish versions of other limit theorems of probability theory (such as the strong law of large numbers, law of the iterated logarithm, and central limit theorem) for predicting several steps ahead, perhaps with an increasing prediction horizon in the case of strong limit theorems;
- replace the assumption of a limited impact horizon in decision problems by softer assumptions of decaying impact.

References

[1] Paul Embrechts and Giovanni Puccetti. Bounds for functions of dependent risks. *Finance and Stochastics*, 10:341–352, 2006.

[2] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, New York, third edition, 1968.

[3] Robert J. Kunsch and Daniel Rudolf. Optimal confidence for Monte Carlo integration of smooth functions. *Advances in Computational Mathematics*, 45:3095–3122, 2019.

[4] Jiří Matoušek and Jan Vondrák. The probabilistic method. Available (in August 2025) on the web, 2008.

[5] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, Cham, second edition, 2022.

[6] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49:1736–1754, 2021.

Acknowledgments

Many thanks to Glenn Shafer for his help and encouragement.