Power and limitations of conformal martingales

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #24

June 2, 2019

Project web site:
http://alrw.net

Abstract

This note poses the problem of investigating the power and limitations of conformal martingales as a means of detecting deviations from randomness. It also gives a crude proposition in this direction and discusses connections between randomness, exchangeability, and conformal martingales.

Contents

1 Introduction

A standard assumption in machine learning has been the assumption that the data are generated in the IID fashion, i.e., independently from the same distribution. This assumption is also known as the assumption of randomness (see, e.g., [8, Section 7.1] and [14]). In this note we will be interested in testing this assumption.

Conformal martingales have been used as a means of detecting deviations from randomness both in theoretical work (see, e.g., [14, Section 7.1], [2], [1]) and in practice (in the framework of the Microsoft Azure module on time series anomaly detection [15]). A natural question is how much we can potentially lose when using conformal martingales as compared with unrestricted testing. This note proposes a formalization of this question. We will see that at a crude scale customary in the algorithmic theory of randomness we do not lose much when restricting our attention to testing randomness with conformal martingales.

Connections with the algorithmic theory of randomness will be explained in Appendix A. The main part of the note will not use the algorithmic notion of randomness; however, as customary in the algorithmic theory of randomness, we will concentrate on the binary case. Besides, we will concentrate on the case of a finite time horizon N.

2 IID probability vs exchangeability probability

Let $\Omega := \{0,1\}^N$ be the set of all binary sequences of length N, and let B_p be product Bernoulli probability measure on Ω with the probability of 1 equal to $p \in [0,1]$: for each $\omega \in \Omega$, $B_p(\{\omega\}) := p^k(1-p)^{N-k}$, where k is the number of 1s in ω. The time horizon $N \in \{1,2,\dots\}$ can be regarded as fixed in this note, apart from the formulas involving $O(\dots)$, which are always uniform in N.

The *upper IID probability* of a set $E \subseteq \Omega$ is defined to be

$$\mathbb{P}^{\mathrm{iid}}(E) := \sup_{p\in[0,1]} B_p(E), \tag{1}$$

and the *upper exchangeability probability* of $E \subseteq \Omega$ is defined to be

$$\mathbb{P}^{\mathrm{exch}}(E) := \sup_{P} P(E), \tag{2}$$

P ranging over the exchangeable probability measures on Ω (a probability measure P on Ω is *exchangeable* if $P(\{\omega\})$ depends only on the number of 1s in ω).

Remark. The lower probabilities corresponding to (1) and (2) are $1-\mathbb{P}^{\mathrm{iid}}(\Omega \setminus E)$ and $1-\mathbb{P}^{\mathrm{exch}}(\Omega \setminus E)$, respectively. In this note we will never need lower probabilities.

The function $\mathbb{P}^{\mathrm{iid}}$ can be used when testing the hypothesis of randomness: if $\mathbb{P}^{\mathrm{iid}}(E)$ is small (say, below 5% or 1%) and the observed sequence ω is in E that is chosen in advance, we are entitled to reject the hypothesis that the observations in ω are IID. Similarly, $\mathbb{P}^{\text{exch}}$ can be used when testing the hypothesis of exchangeability.

Proposition 1. *For any $E \subseteq \Omega$,*

$$\mathbb{P}^{\text{iid}}(E) \le \mathbb{P}^{\text{exch}}(E) \le 1.5\sqrt{N}\,\mathbb{P}^{\text{iid}}(E). \tag{3}$$

Proof. The first inequality in (3) follows from each product Bernoulli probability measure on Ω being exchangeable. If E contains either the all-0 sequence $0\dots0$ or the all-1 sequence $1\dots1$, the second inequality in (3) is obvious. If E is empty, it is also obvious. Finally, if E is nonempty and contains neither sequence, we have, for some $k \in \{1,\dots,N-1\}$,

$$\mathbb{P}^{\text{exch}}(E) = \mathbb{P}^{\text{exch}}(E \cap \Omega_k) = \frac{1/\binom{N}{k}}{(k/N)^k(1-k/N)^{N-k}}\,\mathbb{P}^{\text{iid}}(E \cap \Omega_k) \tag{4}$$

$$\le \frac{k!(N-k)!N^N}{N!k^k(N-k)^{N-k}}\,\mathbb{P}^{\text{iid}}(E) \le \sqrt{2\pi}e^{1/6}\sqrt{\frac{k(N-k)}{N}}\,\mathbb{P}^{\text{iid}}(E) \tag{5}$$

$$\le (\sqrt{2\pi}e^{1/6}/2)\sqrt{N}\,\mathbb{P}^{\text{iid}}(E) \le 1.5\sqrt{N}\,\mathbb{P}^{\text{iid}}(E), \tag{6}$$

where Ω_k is the set of all sequences in Ω containing k 1s. The first equality in (4) follows from each exchangeable probability measure on Ω being a convex mixture of the uniform probability measures on Ω_k, $k = 0,\dots,N$. The second equality in (4) follows from the maximum of $B_p(\{\omega\})$ over $p \in [0,1]$ being attained at $p = k/N$. The first inequality in (5) is equivalent to the obvious $\mathbb{P}^{\text{iid}}(E \cap \Omega_k) \le \mathbb{P}^{\text{iid}}(E)$. The second inequality in (5) follows from Stirling's formula

$$n! = \sqrt{2\pi}n^{n+1/2}e^{-n}e^{r_n}, \quad 0 < r_n < \frac{1}{12n},$$

valid for $n \in \{1,2\dots\}$; see, e.g., [12], where it is also shown that $r_n > \frac{1}{12n+1}$. The first inequality in (6) follows from $\max_{p\in[0,1]} p(1-p) = 1/4$. ☐

Kolmogorov's [5, 6] implicit interpretation of (3) was that $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{exch}}$ are close; on the log scale we have

$$-\log \mathbb{P}^{\text{iid}}(E) = -\log \mathbb{P}^{\text{exch}}(E) + O(\log N),$$

whereas typical values of $-\log \mathbb{P}^{\text{iid}}(E)$ and $-\log \mathbb{P}^{\text{exch}}(E)$ have the order of magnitude N for small (but non-zero) $|E|$. See Appendix A for further details.

3 Conformal probability

In this section we will define the upper conformal probability $\mathbb{P}^{\text{conf}}$, an analogue of $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{exch}}$ for testing using conformal martingales. To define it, we first need to give basic definitions of conformal prediction in the binary case; in this case the definitions simplify greatly.

A *(binary) conformity measure* is a function $A : (n,k) \mapsto A(n,k) \in \{0, 0.5, 1\}$ (automatically measurable) whose domain consists of pairs (n,k) with $n \in \{1,2,\dots\}$ and $k \in \{0,\dots,n\}$ and whose codomain is $\{0, 0.5, 1\}$. Intuitively, $A(n,k)$ tells us whether 1 or 0 is a more typical element of a binary sequence of length n containing k 1s; $A(n,k) = 0.5$ means that 0 and 1 are equally typical. The *p-value* p_n generated by A after being fed with a binary sequence $\omega = (z_1,\dots,z_n) \in \{0,1\}^*$ is

$$p_n = p_n(\omega, \tau_n) := \begin{cases} \tau_n \frac{k}{n} & \text{if } A(n,k) = 0 \text{ and } z_n = 1 \\ \frac{k}{n} + \tau_n \frac{n-k}{n} & \text{if } A(n,k) = 0 \text{ and } z_n = 0 \\ \tau_n & \text{if } A(n,k) = 0.5 \\ \tau_n \frac{n-k}{n} & \text{if } A(n,k) = 1 \text{ and } z_n = 0 \\ \frac{n-k}{n} + \tau_n \frac{k}{n} & \text{if } A(n,k) = 1 \text{ and } z_n = 1, \end{cases} \tag{7}$$

where k is the number of 1s in ω and τ_n is a random number distributed uniformly on the interval $[0,1]$. The standard property of validity for conformal prediction is that the p-values $p_1, p_2, \dots$ are IID and distributed uniformly on $[0,1]$ provided $z_1, z_2, \dots$ are IID, $\tau_1, \tau_2, \dots$ are IID and distributed uniformly on $[0,1]$, and the sequences $z_1, z_2, \dots$ and $\tau_1, \tau_2, \dots$ are independent (see, e.g., [14, Proposition 2.8]).

Perhaps the most natural binary conformity measure is

$$A(n,k) := \begin{cases} 0 & \text{if } k < n/2 \\ 0.5 & \text{if } k = n/2 \\ 1 & \text{if } k > n/2. \end{cases}$$

Therefore, it is interesting that this note only needs the simplest conformity measures, constants; namely, we will use the conformity measure A that is an identical zero: $A(n,k) = 0$ for all n and k.

The next definition is a modification of the definition of "betting functions" in [1]. A *betting martingale* is a measurable function $F : [0,1]^* \to [0,\infty]$ such that, for each sequence $(p_1,\dots,p_{n-1})$, $n \geq 1$, we have

$$\int_0^1 F(p_1,\dots,p_{n-1},p)\,\mathrm{d}p = F(p_1,\dots,p_{n-1});$$

notice that betting martingales are required to be nonnegative. A *nonnegative conformal martingale* is any function $S : (\{0,1\} \times [0,1])^* \to [0,\infty]$ such that, for some conformity measure A and betting martingale F, for all $m \in \{0,1,\dots\}$, $\omega \in \{0,1\}^m$, and $\theta \in [0,1]^m$,

$$S(\omega,\theta) = F(p_1,\dots,p_m)$$

where p_n, $n = 1,\dots,m$, is the p-value computed by (7) from the conformity measure A, the number k of 1s in the prefix of ω of length n, and the nth element τ_n of the sequence $\theta = (\tau_1,\dots,\tau_m) \in [0,1]^m$.

We will define a simple version of upper conformal probability sufficient for our current purpose; there are other natural definitions. The *upper conformal probability* of $E \subseteq \Omega$ is

$$\mathbb{P}^{\mathrm{conf}}(E) := \inf\{S(\Box) \mid \forall \omega \in E : S(\omega, \theta) \geq 1 \text{ a.s.}\}, \tag{8}$$

where S ranges over the nonnegative conformal martingales, $\Box$ is the empty sequence, and "a.s." refers to the uniform probability measure over $\theta \in [0,1]^N$.

The following proposition shows that $\mathbb{P}^{\mathrm{iid}}$ and $\mathbb{P}^{\mathrm{conf}}$ are close, in the sense similar to the closeness of $\mathbb{P}^{\mathrm{iid}}$ and $\mathbb{P}^{\mathrm{exch}}$ asserted in Proposition 1.

Proposition 2. *For any* $E \subseteq \Omega$,

$$\mathbb{P}^{\mathrm{iid}}(E) \leq \mathbb{P}^{\mathrm{conf}}(E) \leq N\, \mathbb{P}^{\mathrm{exch}}(E). \tag{9}$$

Proposition 2 says that, at our crude scale, lack of exchangeability can be detected using conformal martingales. Namely, given a critical region E of a very small size $\epsilon := \mathbb{P}^{\mathrm{exch}}(E)$, we can construct a nonnegative conformal martingale with initial capital $N\epsilon$ that attains capital of 1 when E happens.

In the rest of this section we will check Proposition 2. The following lemma asserts the left inequality in (9) (but in fact proves a stronger statement).

Lemma 1. *For any* $E \subseteq \Omega$, $\mathbb{P}^{\mathrm{iid}}(E) \leq \mathbb{P}^{\mathrm{conf}}(E)$.

Proof. We will check that the statement of the lemma remains true if the right-hand side of (8) is replaced by

$$\inf\{S(\Box) \mid \forall \omega \in E : \mathbb{E}_\theta\, S(\omega, \theta) \geq 1\}, \tag{10}$$

where the $\mathbb{E}_\theta$ refers to the randomness in θ (distributed uniformly in $[0,1]^N$). Let $S_N : \Omega \times [0,1]^N \to [0,\infty]$ be the function mapping each $(\omega,\theta) \in \Omega \times [0,1]^N$ to $S(\omega,\theta)$. It suffices to prove that, for each $E \subseteq \Omega$, each $p \in [0,1]$, and each nonnegative conformal martingale S such that $\mathbb{E}\, S_N \geq 1_E$, we have $B_p(E) \leq S(\Box)$. This follows from the property of validity of conformal martingales:

$$B_p(E) = \mathbb{E}_{B_p}(1_E) \leq \mathbb{E}_{B_p}\, \mathbb{E}\, S_N = S(\Box). \qquad \Box$$

It remains to check the right inequality in (9).

Lemma 2. *For any* $E \subseteq \Omega$,

$$\mathbb{P}^{\mathrm{conf}}(E) \leq N\, \mathbb{P}^{\mathrm{exch}}(E). \tag{11}$$

Proof. Let us first check the second inequality in

$$\mathbb{P}^{\mathrm{iid}}(\{\omega\}) = \frac{k^k (N-k)^{N-k}}{N^N} \leq \mathbb{P}^{\mathrm{conf}}(\{\omega\}) \leq \frac{k!(N-k)!}{N!} = \mathbb{P}^{\mathrm{exch}}(\{\omega\}), \tag{12}$$

where $k \in \{0,\ldots,N\}$ and $\omega \in \Omega$ contains k 1s (all other statements in (12) were established in Proposition 1 and its proof and Lemma 1; they will not be used in the rest of this proof and are given only for symmetry).

To check the second inequality in (12), let $\omega = (z_1, \ldots, z_N)$ be the representation of ω as a sequence of bits. Consider the nonnegative conformal martingale S_ω obtained from the conformity measure $A := 0$ and a betting martingale F such that $F(\Box) = 1/\binom{N}{k}$ and

$$\frac{F(p_1, \ldots, p_{n-1}, p_n)}{F(p_1, \ldots, p_{n-1})} := \begin{cases} \frac{n}{k_n} & \text{if } p_n \leq k_n/n \text{ and } z_n = 1 \\ \frac{n}{n-k_n} & \text{if } p_n \geq k_n/n \text{ and } z_n = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $n = 1, \ldots, N$ and k_n is the number of 1s in ω observed so far,

$$k_n := |\{j \in \{1, \ldots, n\} \mid z_j = 1\}|;$$

in particular, $k_N = k$. (Intuitively, the conformal martingale gambles recklessly on the nth observation being z_n.) If the actual sequence of observations happens to be ω, on step n the value of the martingale is multiplied, a.s., by the fraction whose numerator is n and whose denominator is the number of bits z_n observed so far. The product of all these fractions over $n = 1, \ldots, N$ will have $N!$ as its numerator and $k!(N-k)!$ as its denominator. This conformal martingale is almost deterministic, in the sense of not depending on τ_n provided $\tau_n \notin \{0, 1\}$, and its final value on ω is, a.s.,

$$\frac{1}{\binom{N}{k}} \frac{N!}{k!(N-k)!} = 1.$$

To generalize (11) from singletons to arbitrary $E \subseteq \Omega$, notice that a linear combination of conformal martingales S_ω is again a conformal martingale, since they involve the same conformity measure and betting martingales can be combined. Fix $E \subseteq \Omega$ and remember that Ω_k is the set of all sequences in Ω containing k 1s. Represent E as the disjoint union

$$E = \bigcup_{k=0}^{N} E_k, \quad E_k \subseteq \Omega_k,$$

and let U_k be the uniform probability measure on Ω_k. We then have

$$\begin{aligned} \mathbb{P}^{\mathrm{conf}}(E) \leq \sum_{k=0}^{N} \sum_{\omega \in \Omega_k} \mathbb{P}^{\mathrm{conf}}(\{\omega\}) &\leq \sum_{k=0}^{N} \sum_{\omega \in \Omega_k} \mathbb{P}^{\mathrm{exch}}(\{\omega\}) \\ &= \sum_{k=0}^{N} U_k(E_k) \leq N \max_{k=0}^{N} U_k(E_k) = N\, \mathbb{P}^{\mathrm{exch}}(E), \end{aligned}$$

where the last inequality holds when, e.g., E does not contain the all-0 sequence $0 \ldots 0 \in \Omega$. If E does contain the all-0 sequence, it is still true that

$$\mathbb{P}^{\mathrm{conf}}(E) \leq 1 \leq N = N\, \mathbb{P}^{\mathrm{exch}}(E). \qquad \Box$$

4 Conclusion

Propositions 1 and 2 say that IID, exchangeability, and conformal upper probabilities are close, but the accuracy of these statements is very low and far from meaningful in practice. The most obvious direction of research is to obtain more accurate results (an example related to Proposition 1 will be given in Appendix A). It would be ideal to establish exact bounds on upper conformal probability in terms of upper IID probability and upper exchangeability probability. The most natural definition of upper conformal probability in this context would involve randomness in a more substantial way than our official definition (8) does (cf. (10)).

References

[1] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 4, April 2012. Conference version: ICML 2012.

[2] Shen-Shyang Ho and Harry Wechsler. A martingale framework for detecting changes in data streams by testing exchangeability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:2113–2127, 2010.

[3] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. English translation: *Foundations of the Theory of Probability*. Chelsea, New York, 1950.

[4] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965.

[5] Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions of Information Theory*, IT-14:662–664, 1968.

[6] Andrei N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38:29–40, 1983.

[7] Andrei N. Kolmogorov and Vladimir A. Uspensky. Algorithms and randomness. *Theory of Probability and Its Applications*, 32:389–412, 1987.

[8] Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.

[9] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

[10] Per Martin-Löf. Personal communication, February 2005.

[11] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[12] Herbert Robbins. A remark on Stirling's formula. *American Mathematical Monthly*, 62:26–29, 1955.

[13] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Russian original: О понятии бернуллиевости. Another English translation with proofs: On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 15.

[14] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[15] Xiao Zhang, Peter Lu, Josée Martens, Gary Ericson, and Kent Sharkey. *Time Series Anomaly Detection module in Microsoft Azure*. Microsoft, Seattle, WA, May 2019. Online documentation.

A Connections with the algorithmic theory of randomness

Propositions 1 and 2 are very crude, and Section 4 sets the task of obtaining more accurate result. This appendix explains connections of Proposition 1 with the algorithmic theory of randomness and gives references to some more precise results. It will assume knowledge of some basic notions of that theory.

The notion of randomness has been at the centre of discussions of the foundations of probability for at least 100 years, since Richard von Mises's 1919 article [11]. For von Mises, random sequences (*collectives* in his terminology) were the foundation for probability theory and statistics, and other notions, such as probability, were defined in terms of collectives.

Random sequences have been eclipsed in the foundations of mathematical probability theory by measure since Kolmogorov's 1933 *Grundbegriffe* [3]. In the 1960s Kolmogorov started revival of the interest in random sequences, believing that they are important for understanding the applications of probability theory and statistics. He mainly concentrated on binary sequences (as a simple starting point), in which context he referred to them as *Bernoulli sequences*. Kolmogorov's main publications on the algorithmic theory of randomness were [4, 5, 6]. He was also a co-author of [7], which was based on his ideas and publications, although he did not see the final version of that paper [7, Introduction].

Kolmogorov's notion of randomness for an element ω of a simple finite set M was that $K(\omega) \approx -\log|M|$, where K is Kolmogorov complexity and log is binary log (see [4, Section 4]). Martin-Löf [10] modified this requirement to $K(\omega \mid M) \approx -\log|M|$. In his 1968 paper [5, Section 2] Kolmogorov gave his alternative formalization of von Mises's random sequences, with a reference to Martin-Löf: namely, Kolmogorov said that a binary sequence ω of length N

containing k 1s is *Bernoulli* if

$$K(\omega \mid k, N) \approx \log \binom{N}{k}.$$

It is natural to call the difference

$$d^{\text{exch}}(\omega) := \log \binom{N}{k} - K(\omega \mid k, N) \tag{13}$$

the *exchangeability deficiency* of ω (in terminology close to that of [7]). Being Bernoulli in the sense of Kolmogorov does not fully reflect the intuition of being a plausible outcome of a sequence of N tosses of a possibly biased coin; this intuition is better captured by

$$d^{\text{iid}}(\omega) := \inf_{p \in [0,1]} \left(-\log B_p(\omega) - K(\omega \mid p, N) \right), \tag{14}$$

which we call the *IID deficiency* of ω, being small.

Definitions (13) and (14) can be restated in terms of Martin-Löf's [9] more standard approach using nested families of critical regions. This restatement in combination with Proposition 1 immediately implies

$$d^{\text{exch}}(\omega) - O(1) \le d^{\text{iid}}(\omega) \le d^{\text{exch}}(\omega) + \frac{1}{2} \log N + O(1). \tag{15}$$

In fact, we can interpret (15) as the algorithmic version of Proposition 1. Kolmogorov regarded the coincidence to within log as close enough, at least for some purposes: cf. the last two paragraphs of [5]; therefore, he preferred the simpler definition $d^{\text{exch}}(\omega) \approx 0$ of ω being a Bernoulli sequence.

The difference between natural versions of (13) and (14) is explored in [13, Theorems 1 and 2]. Theorem 1 shows that

$$D^{\text{iid}}(\omega) - \left(\log \binom{N}{k} - KP(\omega \mid N, k, D^{\text{bin}}(k)) \right) = D^{\text{bin}}(k) + O(1),$$

where k is the number of 1s in ω, KP is prefix complexity, D^{iid} is the analogue of d^{iid} using prefix instead of Kolmogorov complexity, and $D^{\text{bin}}(k)$ is the *prefix binomial deficiency* of k defined by

$$D^{\text{bin}}(k) := \inf_{p \in [0,1]} \left(-\log \text{bin}_p(k) - KP(k \mid p, N) \right),$$

bin_p being the binomial probability distribution on $\{0, \dots, N\}$ with parameter p. Theorem 2 characterizes $D^{\text{bin}}(k)$ in terms of prefix complexity, showing that it can be as large as $\frac{1}{2} \log N + O(1)$. These results can be roughly summarized as: for a binary sequence ω to be IID, it needs to be exchangeable and the number k of 1s in it needs to be binomial.

It would be interesting to state Theorems 1 and 2 of [13] without using randomness deficiency, in a form close to Proposition 1. It might also be possible to obtain similar analogues to Proposition 2.