Exchangeability and randomness for infinite and finite sequences

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 #45

First posted December 16, 2025. Last revised January 19, 2026.

Project web site:
http://alrw.net

Abstract

Randomness (in the sense of being generated in an IID fashion) and exchangeability are standard assumptions in nonparametric statistics and machine learning, and relations between them have been a popular topic of research. This short paper draws the reader's attention to the fact that, while for infinite sequences of observations the two assumptions are almost indistinguishable, the difference between them becomes very significant for finite sequences of a given length.

Contents

1 Introduction

In this paper we will discuss two important assumptions about sequences of observations, exchangeability and randomness, using the word "randomness" in a somewhat old-fashioned sense of the individual observations being independent and identically distributed (following [20, Chap. 7], which used the standard terminology for its time, and [28]).

The relationship between exchangeability and randomness is very different in the cases of infinite sequences and finite sequences. In the former case, there is hardly any difference between the two assumptions. But in the latter, the difference may be vast. The study of this relationship has a long history, which will also be briefly reviewed.

We start in Sect. 2 with a discussion of de Finetti's theorem, which was later greatly generalized by Hewitt and Savage and other people. One implication of de Finetti's theorem is that, for a wide range of observation spaces, there is no difference between the assumptions of exchangeability and randomness.

In Sect. 3 we move on to the case of finite sequences of observations of a given length. The problem of relation between exchangeability and randomness in this case was implicitly posed by Kolmogorov [17] in his work on the frequentist foundations of probability. In the context of the algorithmic theory of randomness, he simply defined randomness as exchangeability for binary sequences (in which case the difference between the two assumptions is much less significant, as discussed in Sect. 4.2). Precise difference for a natural alternative definition of randomness was explored in work done under his supervision [26]. In Sect. 3.1 we will see a simple non-binary example where the difference between exchangeability and randomness is very substantial.

According to Kolmogorov, and it is difficult to argue with his point of view, infinite sequences are empirically vacuous; we never observe them in reality. Therefore, he always insisted on either studying finite sequences or at least keeping them in mind. In fact, de Finetti's results can also be stated in terms of finite sequences, provided we consider sequences of different lengths. In this paper we will briefly discuss two modes of statistical hypothesis testing involving finite sequences, online and batch; the former corresponds to de Finetti's setting and the latter to Kolmogorov's.

In Sect. 4 we check that the gap between exchangeability and randomness described in Sect. 3 is the widest possible in some sense. In its first part, Sect. 4.1, we discuss a simpler and cleaner mathematical result that holds for an infinite observation space. Then Sect. 4.2 is devoted to more complicated and perhaps less practically relevant results about finite observation spaces. However, these results cover Kolmogorov's binary case. And we will see that exchangeability and randomness are asymptotically close in this case, at least according to a relaxed standard proposed and sometimes used by Kolmogorov.

This paper has little novelty, and all of its mathematical results are very simple. But the reader might find the scale of the difference between exchangeability and randomness for infinite observation spaces surprising.

2 De Finetti's theorem

Suppose we observe a data sequence $z_1, z_2, \cdots \in \mathbf{Z}$ consisting of observations $z_n \in \mathbf{Z}$ that are elements of some measurable space $\mathbf{Z}$, the *observation space*. The *assumption of randomness* is standard in machine learning: the data is coming from the product probability measure Q^∞ for some $Q \in \mathfrak{P}(\mathbf{Z})$, where $\mathfrak{P}(Z)$ stands for the measurable space of all probability measures on a measurable space Z (with the σ-algebra generated by the mappings $Q \mapsto Q(A)$, A being an event in Z).

Remark 1. Effects of various distribution shifts have also been widely studied, but in this paper we concentrate on the basic IID case. In machine learning we often have observations $z = (x, y)$ consisting of an object x and its label y, but we do not insist on this.

The more general *assumption of exchangeability* is that the data is coming from an *exchangeable* probability measure $R \in \mathfrak{P}(\mathbf{Z}^\infty)$, i.e., a probability measure that is invariant w.r. to swapping any pair of observations. The topic of this section is the closeness of the two assumptions for an infinite, or at least potentially infinite, data sequences and assuming that $\mathbf{Z}$ is a *Borel space*, i.e., a measurable space that is isomorphic to a Borel subset of $\mathbb{R}$.

By de Finetti's classical theorem each exchangeable probability measure R on $\mathbb{R}^\infty$ is a mixture of product distributions:

$$R = \int Q^\infty \mu(\mathrm{d}Q) \tag{1}$$

for some $\mu \in \mathfrak{P}(\mathfrak{P}(\mathbb{R}))$. This was established by de Finetti [4, Chap. 4]. Hewitt and Savage [11, Theorem 7.3 and its discussion later in Sect. 7] note that we can trivially replace $\mathbb{R}$ by any Borel space $\mathbf{Z}$ and point out that, to the best of their knowledge, every measurable space known to have importance in applied science is Borel. For example, every Polish space (complete separable metric space with its Borel σ-algebra) is Borel. On the other hand, there exists a separable metric space with its Borel σ-algebra for which (1) can be violated [8]. Let us assume in the rest of this paper that the observation space $\mathbf{Z}$ is Borel.

It is well known that de Finetti's theorem fails if we simply replace ∞ by a finite N in (1); see, e.g., [3, Sect. 4.7.1] (and [6, Sect. 1] for a further discussion of the extent to which de Finetti's theorem can fail for $N = 2$ and $\mathbf{Z} = \{0, 1\}$).

De Finetti's theorem plays an important role in the foundations of Bayesian statistics; see, e.g., [3, Chap. 4]. But its implication in our context is that the assumptions of exchangeability and randomness are equivalent: if a testing procedure rejects one of these assumptions, it rejects the other as well. We will formalize this statement in three different ways.

The simplest way of statistical hypothesis testing is based on critical regions. To test a composite null hypothesis $\mathcal{H}$ (a set of probability measures) at a significance level $\epsilon \in (0, 1)$ (such as 1% or 5%), we choose a *critical region* A at level ϵ, meaning an event satisfying $R(A) \le \epsilon$ for all $R \in \mathcal{H}$. The hypothesis $\mathcal{H}$ is rejected if we observe A, which must be chosen in advance. Applying

this to the assumption of exchangeability, critical regions $A \subseteq \mathbf{Z}^\infty$ for testing exchangeability at level ϵ are required to satisfy $\mathbb{P}^{\mathrm{X}}(A) \le \epsilon$, where the *upper exchangeability probability* of A is defined by

$$\mathbb{P}^{\mathrm{X}}(A) := \sup_R R(A), \tag{2}$$

R ranging over the exchangeable probability measures on $\mathbf{Z}^\infty$. Similarly, critical regions $A \subseteq \mathbf{Z}^\infty$ for testing randomness at level ϵ are required to satisfy $\mathbb{P}^{\mathrm{R}}(A) \le \epsilon$, where the *upper randomness probability* of A is defined as

$$\mathbb{P}^{\mathrm{R}}(A) := \sup_{Q \in \mathfrak{P}(\mathbf{Z})} Q^\infty(A). \tag{3}$$

By de Finetti's theorem, $\mathbb{P}^{\mathrm{X}}$ and $\mathbb{P}^{\mathrm{R}}$ coincide:

- since every product measure Q^∞ is exchangeable, $\mathbb{P}^{\mathrm{R}} \le \mathbb{P}^{\mathrm{X}}$;
- on the other hand, since each exchangeable probability measure R has a representation (1), we have, for each $\alpha > 0$ and each event A satisfying $\mathbb{P}^{\mathrm{R}}(A) \le \alpha$,

$$\mathbb{P}^{\mathrm{X}}(A) = \sup_R R(A) \le \sup_{\mu, Q} \int Q^\infty(A) \mu(\mathrm{d}Q) \le \alpha; \tag{4}$$

  therefore, $\mathbb{P}^{\mathrm{X}} \le \mathbb{P}^{\mathrm{R}}$.

We can see that there are exactly the same critical regions at each significance level under exchangeability and under randomness.

One manifestation of the coincidence of the critical regions under exchangeability and randomness is that the two assumptions will produce identical prediction sets

$$\Gamma(z_1, \dots, z_n) := \left\{ (z_{n+1}, z_{n+2}, \dots) : (z_1, \dots, z_n, z_{n+1}, z_{n+2}, \dots) \notin A \right\} \tag{5}$$

for the future observations after observing $z_1, \dots, z_n$, where A is a critical region at some significance level ϵ. Under both exchangeability and randomness, the coverage probability of the prediction set (5) will be at least $1 - \epsilon$.

There are two popular generalizations of critical regions, p-variables and e-variables, and both also produce identical results under exchangeability and randomness. Let us first check this for p-variables. According to the general definition (see, e.g., [22, Definition 1.2]), an *exchangeability p-variable* is a random variable $P : \mathbf{Z}^\infty \to [0, 1]$ such that, for all $\epsilon \in (0, 1)$,

$$\mathbb{P}^{\mathrm{X}}(P \le \epsilon) \le \epsilon. \tag{6}$$

Similarly, a *randomness p-variable* is a random variable $P : \mathbf{Z}^\infty \to [0, 1]$ such that, for all $\epsilon \in (0, 1)$,

$$\mathbb{P}^{\mathrm{R}}(P \le \epsilon) \le \epsilon. \tag{7}$$

Since $\mathbb{P}^{\mathrm{X}} = \mathbb{P}^{\mathrm{R}}$, the classes of exchangeability and randomness p-variables coincide.

Again according to the general definition [22, Definition 1.2] (see also [10, (1)]), an *exchangeability e-variable* is a measurable function $F : \mathbf{Z}^\infty \to [0, \infty]$ such that $\mathbb{E}^{\mathrm{X}}(F) \le 1$, where

$$\mathbb{E}^{\mathrm{X}}(F) := \sup_R \int F \, \mathrm{d}R, \tag{8}$$

R ranging over the exchangeable probability measures on $\mathbf{Z}^\infty$. And a *randomness e-variable* is a measurable function $F : \mathbf{Z}^\infty \to [0, \infty]$ such that $\mathbb{E}^{\mathrm{R}}(F) \le 1$, where

$$\mathbb{E}^{\mathrm{R}}(F) := \sup_{Q \in \mathfrak{P}(\mathbf{Z})} \int F \, \mathrm{d}Q^\infty. \tag{9}$$

Let us check that $\mathbb{E}^{\mathrm{X}} = \mathbb{E}^{\mathrm{R}}$. Since $\mathbb{E}^{\mathrm{R}} \le \mathbb{E}^{\mathrm{X}}$ is obvious, we just need to check $\mathbb{E}^{\mathrm{X}} \le \mathbb{E}^{\mathrm{R}}$. Generalizing (4), we have, for each $\alpha > 0$, each measurable function $F : \mathbf{Z}^\infty \to [0, \infty]$ satisfying $\mathbb{E}^{\mathrm{R}}(F) \le \alpha$, and each $\delta > 0$,

$$\mathbb{E}^{\mathrm{X}}(F) \le \int F \, \mathrm{d}R + \delta = \iint F \, \mathrm{d}Q^\infty \mu(\mathrm{d}Q) + \delta \le \alpha + \delta, \tag{10}$$

and so indeed $\mathbb{E}^{\mathrm{X}} \le \mathbb{E}^{\mathrm{R}}$. The first inequality in (10) holds for some exchangeable probability measure R and the second for some $\mu \in \mathfrak{P}(\mathfrak{P}(\mathbf{Z}))$ according to (1). As $\mathbb{E}^{\mathrm{X}} = \mathbb{E}^{\mathrm{R}}$, the classes of exchangeability and randomness e-variables coincide.

Whatever testing method out of the three we use, we have the same options for rejecting exchangeability or randomness. The empirical contents of the two assumptions may be said to coincide, under our assumption of $\mathbf{Z}$ being a Borel space.

3 Finite sequences of observations

In the previous section we saw that, by de Finetti's theorem, the difference between the assumptions of exchangeability and randomness disappears. However, this is a statement about infinite sequences, which we can never observe in reality. Can we say something similar for finite sequences of observations?

3.1 Batch setting

We start our discussion of finite sequences of observations from the simplest setting in which we fix the number N of observations and only consider the sequences of length N, $\mathbf{Z}^N$. We call this the *batch setting*. In this case a chasm between exchangeability and randomness opens up. Consider the following very simple critical region A. The N observations are all in the set $\{1, \dots, N\}$ and are all different. Under exchangeability, the event A is perfectly possible: its probability is 1 under some exchangeable probability measure. Defining $\mathbb{P}^{\mathrm{X}}$, $\mathbb{P}^{\mathrm{R}}$, $\mathbb{E}^{\mathrm{X}}$, and $\mathbb{E}^{\mathrm{R}}$ as in the previous section (see (2), (3), (8), and (9)) but replacing

$\mathbf{Z}^\infty$ with $\mathbf{Z}^N$ and Q^∞ with Q^N, we can see that $\mathbb{P}^{\mathrm{X}}(A) = 1$. The maximum probability of A under any product measure Q^N is

$$\frac{N}{N}\frac{N-1}{N}\cdots\frac{1}{N} = \frac{N!}{N^N} \sim \sqrt{2\pi N}\mathrm{e}^{-N}; \tag{11}$$

therefore, it shrinks exponentially fast as N grows. Indeed, the maximum of $Q^N(A)$ is achieved for the Q concentrated on $\{1,\dots,N\}$ and uniformly distributed on this set. Applying Stirling's formula in the form of [23] gives

$$\mathbb{P}^{\mathrm{R}}(A) < 3\sqrt{N}\mathrm{e}^{-N} < 1 = \mathbb{P}^{\mathrm{X}}(A) \tag{12}$$

(the second inequality "<" assumes $N > 1$). For example, suppose we are interested in significance level $\epsilon := 10^{-k}$ for $k \geq 2$ (such as $k = 2$ for high statistical significance). Solving $3\sqrt{N}\mathrm{e}^{-N} \leq 10^{-k}$, we obtain

$$\mathbb{P}^{\mathrm{R}}(A) < 10^{-k} < 1 = \mathbb{P}^{\mathrm{X}}(A)$$

provided $N \geq 3k + 1$. For example, an outcome that is perfectly possible under exchangeability ($\mathbb{P}^{\mathrm{X}}(A) = 1$) becomes highly statistically significant under randomness for $N = 7$. The much stricter significance level of "5 sigma", approximately $1/(3 \times 10^6)$, used for announcing discoveries in particle physics [1] is met starting from $N = 22$.

Remark 2. A convenient, albeit asymptotically much cruder, version of the inequality (12) is

$$\mathbb{P}^{\mathrm{R}}(A) \leq 2^{-N+1} \leq 1 = \mathbb{P}^{\mathrm{X}}(A).$$

Remark 3. The difference between exchangeability and randomness is also manifested by the fact that exchangeability is easy to achieve in practice: we can just permute randomly our data sequence. However, the resulting sequence may be very far from being IID.

Similarly to (5), we can invert exchangeability and randomness critical regions in the batch mode and use them for one-step-ahead prediction (whereas prediction in (5) was infinitely many steps ahead). Given an observed sequence $z_1,\dots,z_n$, we output

$$\Gamma(z_1,\dots,z_n) := \{z_{n+1} : (z_1, z_2, \dots, z_{n+1}) \notin A\}$$

as our prediction set for the next observation, where A is a critical region in $\mathbf{Z}^N$ with $N := n + 1$; this ensures a coverage probability of at least $1 - \epsilon$, where ϵ is the significance level used in A. Under randomness, we can produce non-vacuous (i.e., different from $\mathbf{Z}$) prediction sets at significance level $3\sqrt{N}\mathrm{e}^{-N}$ even in situations where no non-vacuous prediction sets are possible under exchangeability at any non-trivial significance level ϵ (i.e., at any $\epsilon < 1$). It is instructive to compare this with conformal prediction, where non-vacuous prediction sets are only possible at much larger significance levels of at least $1/N$ [28, Sect. 11.4.4].

3.2 Kolmogorov's and Martin-Löf's work on Bernoulli sequences

The assumptions of randomness and exchangeability, in different guises, have also played important roles in the foundations of frequentist statistics. The standard measure-theoretic foundations of probability were put forward in Kolmogorov's *Grundbegriffe* [14], but Kolmogorov did not believe that they were sufficient for applications of probability. As he pointed out in the *Grundbegriffe* [14, footnote 4], in his frequentist analysis of the applications of probability he was following Richard von Mises. However, a big difference between Kolmogorov's and von Mises's approaches was that von Mises's was based on infinite sequences, whereas Kolmogorov believed that infinite sequences, being empirically non-existent, had no place in discussions of real-world applications of probability. Interestingly, because of this he even objected against publication in *Russian Mathematical Surveys* (a journal that he edited at the time) of a planned paper about infinite random sequences by his student Uspensky and close collaborators Shen and Semenov; see Kolmogorov's letter to Uspensky of June 1983 cited in [24, note 14].

For a long time Kolmogorov believed that no frequentist concept of probability can be developed for finite sequences [15, Sect. 1], but in 1963 he published his first attempt in this direction [15, Sect. 2]. The attempt, however, was "incomplete" (as he characterizes it in [16, Sect. 4]), and he greatly improved on it in 1968 [17, Sect. 2] (this paper is based on his 1967 talk). It appears that the details of Kolmogorov's improved approach first appeared in print in Martin-Löf's 1966 paper [21, Sect. 5].

Both Kolmogorov and Martin-Löf consider binary sequences and define what they call Bernoulli sequences, i.e., sequences that can be plausibly obtained as result of IID observations. While their informal explanations clearly show that they are interested in the assumption of randomness,[1] their formal definitions are about the assumption of exchangeability. Let me give essentially Martin-Löf's definition; I will use slightly different terminology, but my definition will be equivalent to Martin-Löf's. This definition will rely on some basic notions of the theory of algorithms, but it will not be used outside this subsection, and the reader can skip the rest of the subsection without interrupting the flow of ideas.

Let us define exchangeability p-variables and randomness p-variables as in the previous section, by (6) and (7), but replacing $\mathbf{Z}^\infty$ with $\mathbf{Z}^N$ for a finite N. We consider families P_N, $N \in \{1, 2, \dots\}$, of exchangeability p-variables such that $P_N(x)$ is upper semicomputable as function of N and $x \in \mathbf{Z}^N$ (where the upper semicomputability means that, for a computable function f taking rational values, $P_N(x) = \inf_k f(N, x, k)$, k ranging over the natural numbers). There exists a smallest, to within a constant factor, function $(N, x) \mapsto P_N(x)$

[1] This is a relevant quote from Kolmogorov [17, Sect. 2]: "let us consider how we imagine a sequence of zeros and ones appearing as the result of independent trials with probability p of obtaining a one at each trial."

of this kind, and Martin-Löf defines

$$m(x) := -\log_2 P_N(x)$$

for all binary sequences $x \in \{0,1\}^*$, where N is the length of x and $\log_b$ is base-b logarithm. This definition is slightly arbitrary, since $m(x)$ is only defined to within an additive constant, so additive terms of $O(1)$ are typically ignored in the algorithmic theory of randomness.

Another equivalent definition of m is given by Kolmogorov in [17, Sect. 2] in terms of his notion of complexity (and Martin-Löf proves the equivalence in [21, Sect. 5]). A binary sequence x is called a *Bernoulli sequence* if $m(x)$ is small; this is an informal notion, but we can prove mathematical results about the function m (albeit only with the $O(1)$ accuracy).

Replacing the assumption of exchangeability by that of randomness, we get a function that we denote by m^{R} instead of m. (A natural notation for m in view of our notations $\mathbb{P}^{\mathrm{X}}$ and $\mathbb{E}^{\mathrm{X}}$ would have been m^{X}, but m is what Martin-Löf used in [21, Sect. 5].)

These definitions can be adapted verbatim to the case where the observation space $\mathbf{Z}$ is $\{1,2,\dots\}$ rather than $\{0,1\}$. In this case the example in Sect. 3.1 demonstrating (12) shows that there are sequences $x_N \in \mathbf{Z}^N$, $N = 1,2,\dots$, such that

$$m(x_N) = O(1) < N \log_2 \mathrm{e} - \frac{1}{2}\log_2 N + O(1) = m^{\mathrm{R}}(x_N),$$

the inequality holding from some N on; this is stated without proof in [27, Theorem 4]. We can see that the difference between m and m^{R} is very substantial.

3.3 Versions of de Finetti's theorem for finite sequences of observations

De Finetti's theorem is about infinite sequences of observations, similarly to von Mises's frequentist story criticised by Kolmogorov. Does it mean that de Finetti's theorem is empirically irrelevant? Not at all. Whereas von Mises's story may be hopelessly stuck at infinity (to use Shafer's expression [25]), de Finetti's theorem may be applied to finite sequences, albeit not in the batch setting of Sect. 3.1. A popular alternative to the batch setting is the *online setting*, where we do not fix the number of observations in advance and process them sequentially, and then the sequence of observations becomes potentially infinite; therefore, de Finetti's theorem becomes applicable if we assume exchangeability for all those potential observations. For example, in Sect. 6 of [21] Martin-Löf develops a way of testing exchangeability for all finite prefixes of a potentially infinite sequence of observations, which means that he is indeed testing randomness of the overall sequence.

An instructive finite form of de Finetti's theorem was derived by Diaconis and Freedman [7] (with an early version given already at the very end of [11]). Since the assumptions of exchangeability and randomness are so different for

a fixed length N, we have to consider different lengths for imposing exchangeability and for claiming randomness. In the abstract of [6], which states the Diaconis–Freedman result for binary sequences, Diaconis summarizes this result thus: "an exchangeable sequence of length r which can be extended to an exchangeable sequence of length k is almost a mixture of independent experiments, the error going to zero like $1/k$". It is essential that k and r should be different here, ideally $k \gg r$. There have been several recent information-theoretic developments of this idea; see [12, Corollary 1] for a particularly strong result.

Another popular finite version of de Finetti's theorem appears in Dellacherie and Meyer [5, Chap. 5, 52], who credit this result to P. Cartier; see Kerns and Székely [13] for a fuller exposition. Theorem 1.1 in [13] says that the representation (1) holds in the batch setting for any exchangeable probability measure R on $\mathbf{Z}^N$ without any restrictions on the measurable space $\mathbf{Z}$ if we allow μ to be a signed measure of bounded variation. This results appears to be a mathematical curiosity that does not have any implications for statistical hypothesis testing.

Finally, "de Finetti's theorem" is sometimes used in a much wider sense covering representations of exchangeable probability measures as mixtures of probability measures different from product measures, as in [6, Theorem 1]. We do not discuss such results as our main interest is relations between exchangeability and randomness.

4 Tight inequalities

We first consider the simpler case of an infinite observation space $\mathbf{Z}$ and then move on to the slightly messier case of a finite $\mathbf{Z}$. We always assume that all singletons in $\mathbf{Z}$ are measurable; among other things, this will ensure that the effective size of a finite $\mathbf{Z}$ is equal to its cardinality $|\mathbf{Z}|$.

4.1 The case of infinite Z

In this subsection we will check that the example given in the previous section (see (11) and (12)) is the most extreme when the observation space $\mathbf{Z}$ is infinite. Namely, we will see that, for any $\mathbf{Z}$ (finite or infinite) and any event A in $\mathbf{Z}^N$,

$$\mathbb{P}^{\mathrm{X}}(A) \le \frac{N^N}{N!}\,\mathbb{P}^{\mathrm{R}}(A); \tag{13}$$

the example demonstrates that the equality here is attained. The length N is fixed throughout this section.

The equation (13) can be strengthened to the following proposition.

Proposition 4. *For any random variable $F : \mathbf{Z}^N \to [0, \infty)$,*

$$\mathbb{E}^{\mathrm{X}}(F) \le \frac{N^N}{N!}\,\mathbb{E}^{\mathrm{R}}(F). \tag{14}$$

Suppose $|\mathbf{Z}| \geq N$. *The bound is tight and attained on a non-zero indicator function. Moreover, there exists an event* A *in* $\mathbf{Z}^N$ *such that*

$$\mathbb{P}^{\mathrm{R}}(A) = \frac{N!}{N^N} \leq 1 = \mathbb{P}^{\mathrm{X}}(A).$$

Proof. To prove (14), first notice that the left-hand side is the supremum of the averages of F over all orbits, where an *orbit* is defined to be the set of all permutations (not necessarily distinct) of a sequence in $\mathbf{Z}^N$; as a formula,

$$\mathbb{E}^{\mathrm{X}}(F) = \sup_{z_1,\dots,z_N} \bar{F}(z_1,\dots,z_N),$$

where

$$\bar{F}(z_1,\dots,z_N) := \frac{1}{N!} \sum_{\pi\in S_N} F(z_{\pi(1)},\dots,z_{\pi(N)})$$

and S_N stands for the symmetric group of all permutations of $\{1,\dots,N\}$ (this follows from, e.g., [28, Lemma A.3]). Therefore, it suffices to consider only F that are non-zero on one orbit only. Let F be non-zero on the orbit generated by a sequence $z_1,\dots,z_N$ in $\mathbf{Z}^N$ containing K distinct elements of $\mathbf{Z}$ with multiplicities $n_1,\dots,n_K$ (so that $n_1+\dots+n_K = N$). The largest product probability Q^N of an element of this orbit is

$$\prod_{k=1}^{K} \left(\frac{n_k}{N}\right)^{n_k},$$

and, therefore,

$$\begin{aligned}\mathbb{E}^{\mathrm{R}}(F) = \bar{F}(z_1,\dots,z_N)\frac{N!}{n_1!\dots n_K!}\prod_{k=1}^{K}\left(\frac{n_k}{N}\right)^{n_k} \\ = \mathbb{E}^{\mathrm{X}}(F)\frac{N!}{n_1!\dots n_K!}\prod_{k=1}^{K}\left(\frac{n_k}{N}\right)^{n_k} \geq \mathbb{E}^{\mathrm{X}}(F)\frac{N!}{N^N}, \quad (15)\end{aligned}$$

where the inequality follows from $n^n/n! \geq 1$. This completes the proof of (14).

The example in the previous section demonstrates that the bound is tight when $|\mathbf{Z}| \geq N$, since we may assume $\{1,\dots,N\} \subseteq \mathbf{Z}$ without further loss of generality. □

4.2 Smaller observation spaces Z

The case $|\mathbf{Z}| = \infty$, or at least $|\mathbf{Z}| \gg 1$, is probably most relevant in practice (e.g., even in the case of classification problems, the objects to be classified are typically complex). In this section, however, we consider the case of a finite $\mathbf{Z}$ and are mostly interested in a small $|\mathbf{Z}|$. This will allow us to cover, e.g., Kolmogorov's and Martin-Löf's case of binary sequences.

The inequality (14) in Proposition 4 is tight in the case of an infinite $\mathbf{Z}$, which we relaxed in the statement of the proposition to what is actually used in the proof, $|\mathbf{Z}| \geq N$. The following proposition extends (14) to smaller $\mathbf{Z}$.

Proposition 5. *Suppose $K := |\mathbf{Z}| < N$. For any random variable $F : \mathbf{Z}^N \to [0, \infty)$, we have $\mathbb{E}^{\mathrm{X}}(F) \le C\,\mathbb{E}^{\mathrm{R}}(F)$, where*

$$C := \frac{N^N}{N!} \prod_{k=1}^{K} \frac{n_k!}{n_k^{n_k}} \tag{16}$$

and n_k is any balanced split of N into K parts:

$$n_k \in \{\lfloor N/K \rfloor, \lceil N/K \rceil\} \text{ such that } \sum_{k=1}^{K} n_k = N. \tag{17}$$

The factor C is tight; moreover, there exists an event $A \subseteq \mathbf{Z}^N$ such that

$$\mathbb{P}^{\mathrm{R}}(A) = 1/C \le 1 = \mathbb{P}^{\mathrm{X}}(A). \tag{18}$$

Proof. Without loss of generality, we set $\mathbf{Z} := \{1, \dots, K\}$ and proceed as in the proof of Proposition 4. Let n_k be the number of times that $k \in \mathbf{Z}$ occurs in $z_1, \dots, z_N$ (so that now $n_k = 0$ is possible, in which case $n_k^{n_k} := 1$ and $n_k! = 1$). Now instead of the inequality "$\ge$" in (15) we use the convexity of the function $\log(n^n/n!) = n \log n - \log(n!)$ in n (see Lemma 6 below). If the split n_k, $k = 1, \dots, K$, of N is not balanced, we can find n_{k_1} and n_{k_2} such that $n_{k_1} - n_{k_2} \ge 2$. By the convexity, the penultimate expression in the chain (15) cannot increase if we move n_{k_1} and n_{k_2} towards each other by redefining $n_{k_1} := n_{k_1} - 1$ and $n_{k_2} := n_{k_2} + 1$. Repeating this operation we arrive at a balanced split.

An event A satisfying (18) can be chosen as the orbit with the counts n_k for $k \in \mathbf{Z}$ given by (17). □

The following lemma was used in the proof of Proposition 5.

Lemma 6. *The function $n \log n - \log(n!)$ of $n \in \{0, 1, \dots\}$ is convex (and even strictly convex).*

Proof. We are required to prove that the function

$$f(n) := ((n+1)\log(n+1) - \log((n+1)!)) \\ - (n \log n - \log(n!)) = n \log(1 + 1/n)$$

is strictly increasing. Extending it to the nonnegative reals, $f(x) := x \log(1 + 1/x)$, interpreting log as ln, and differentiating gives

$$f'(x) = \log\left(1 + \frac{1}{x}\right) - \frac{1}{x+1}.$$

By the strict concavity of log, we have $\log(1+u) > \frac{u}{1+u}$ for $u > 0$, and substituting $u := 1/x$ gives $f'(x) > 0$ for all $x > 0$. □

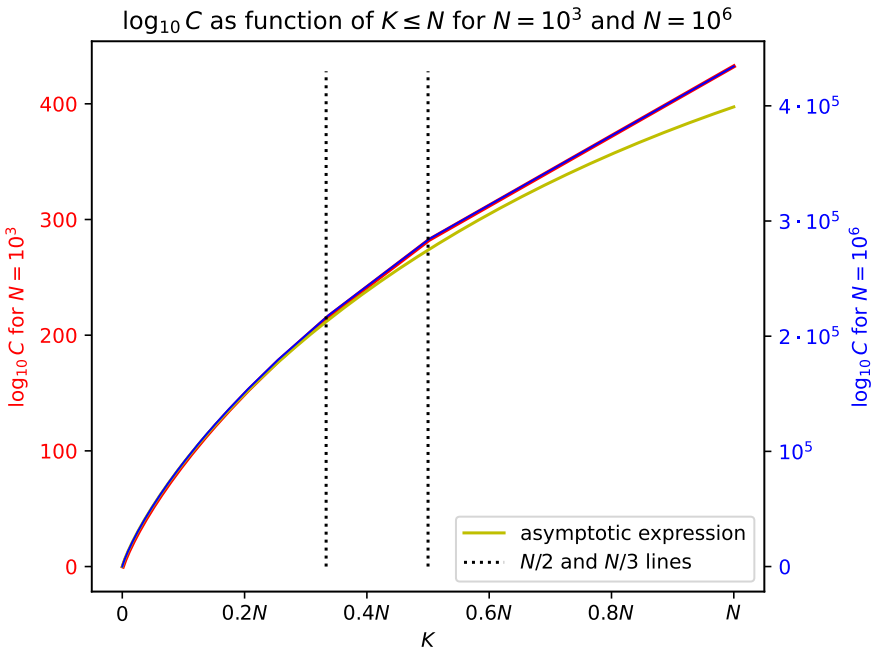

Figure 1: The graphs of $\log_{10} C$ for $N = 10^3$ (thick red line, with some values given on the left) and $N = 10^6$ (thin blue line, with values on the right). The yellow line represents the approximation described in Remark 7.

Figure 1 shows $\log_{10} C$ as a function of $K \in \{1, \ldots, N\}$ for two values of N, 10^3 (red, with the scale of $\log_{10} C$ on the left) and 10^6 (blue, with the scale on the right). Both graphs become horizontal to the right of $K := N$ (not shown in Fig. 1). We can see that the shapes of the two graphs are very similar. The blue line (for 10^6) was drawn after the red one (for 10^3), and the latter is thicker in order to be able to see the small difference between the graphs. When implementing the formula (16) for C, it is convenient to compute the number of $n_k = \lceil N/K \rceil$ in (17) as $N \bmod K$.

The vertical dotted lines in Fig. 1 are drawn through the observation $N/2$ (the right line) and the nearest observation to $N/3$ (the left one). The qualitative behaviour of the red and blue lines, namely the strictly exponential growth (linear on the log scale of Fig. 1) of both graphs between, roughly, $N/2$ and N, $N/3$ and $N/2$, etc., is easy to understand. When K increases by 1 between $N/2$ and N, one of the $n_k = 2$ in (17) gets replaced by two n_k, namely 1 and 1. Therefore, the expression for C given by (16) gets multiplied by $2^2/2! = 2$, and so the slope of the red and blue lines between $N/2$ and N in Fig. 1 is $\log_{10} 2 \approx 0.301$. When K increases by 1 between $N/3$ and $N/2$, a block $(3, 3)$ of two n_k in (17) gets replaced by a block $(2, 2, 2)$ of three n_k. The expression

for C given by (16) gets multiplied by

$$\left(\frac{2!}{2^2}\right)^3 / \left(\frac{3!}{3^3}\right)^2 = 3^4/2^5,$$

and so the slope of the red and blue lines between $N/3$ and $N/2$ in Fig. 1 is $\log_{10}(3^4/2^5) \approx 0.403$. And when K increases by 1 between $N/4$ and $N/3$, a block $(4,4,4)$ of three n_k gets replaced by a block $(3,3,3,3)$ of four n_k. The expression for C gets multiplied by

$$\left(\frac{3!}{3^3}\right)^4 / \left(\frac{4!}{4^4}\right)^3 = 2^{19}/3^{11},$$

and so the slope of the red and blue lines immediately to the left of $N/3$ in Fig. 1 is $\log_{10}(2^{19}/3^{11}) \approx 0.471$. The slope keeps increasing as we move left.

Remark 7. Let us replace the decimal logarithms $\log_{10}$ by natural ln in the graphs shown in Fig. 1 for $N \in \{10^3, 10^6\}$. This will not change the shape of the graphs and will only change the labels on the axes; the upper limit of the range of $\log C$ will now become close to N (this will be checked in the appendix). An interesting function is the limit L of these graphs as $N \to \infty$ with both axes rescaled by dividing by N (so that the slopes remain unchanged). It can be defined as the continuous piecewise-linear function $L : [0,1] \to [0,\infty)$ satisfying $L(0) := 0$ and

$$L'(x) := \ln\left(\left(\frac{n!}{n^n}\right)^{n+1}\left(\frac{(n+1)!}{(n+1)^{n+1}}\right)^{-n}\right) \text{ for all } x \in \left(\frac{1}{n+1}, \frac{1}{n}\right) \tag{19}$$

and for all $n \in \{1,2,\dots\}$. As $x \to 0$,

$$L(x) = -\frac{1}{2}x\ln x + \frac{\ln(2\pi)}{2}x + o(x),$$

and the right-hand side without the "$+ o(x)$" and with $\log_{10}$ in place of ln is shown as the yellow line in Fig. 1. The final value of the approximation $-\frac{1}{2}x\ln x + \frac{\ln(2\pi)}{2}x$ at $x = 1$ is approximately 0.919, which is not so different from $L(1) = 1$. We can see that the slope of $L(x)$ is infinity at $x = 0$.

Another interesting case for a finite $\mathbf{Z}$ is where $K := |\mathbf{Z}|$ is fixed while the number of observations N varies. Applying Stirling's formula to (16) we then obtain

$$C = \Theta(N^{(K-1)/2}). \tag{20}$$

This polynomial growth rate as $N \to \infty$ contrasts with the exponential growth rate for $|\mathbf{Z}| = \infty$. The closeness of $\mathbb{E}^{\mathrm{X}}$ and $\mathbb{E}^{\mathrm{R}}$ to within a polynomial factor (namely, $\Theta(N^{(K-1)/2})$) implies the closeness of e-variables under exchangeability and randomness in the same crude sense; it also implies the closeness of $\mathbb{P}^{\mathrm{X}}$ and $\mathbb{P}^{\mathrm{R}}$, which in turn implies the closeness of p-variables under exchangeability and randomness, in the same sense.

On Kolmogorov's and Martin-Löf's log scale, $N^{(K-1)/2}$ becomes $\frac{K-1}{2}\log N$, and differences of $O(\log N)$ are often ignored in the algorithmic theory of randomness; according to Kolmogorov, "we should not be afraid of logarithms (as well as $O(1)$ terms that we have anyway)".[2] In this sense exchangeability and being IID nearly coincide for finite sequences as well. However, for $|\mathbf{Z}| = \infty$ the difference becomes $\Theta(N)$ on the logarithmic scale, which is the largest possible in Kolmogorov's and Martin-Löf's binary setting.

More recently the log scale for p-values has been advocated by Greenland, who called $-\log_2 p$ the *S-value*, or *p-surprisal*, corresponding to a p-value of p; see, e.g., [9]. There have been no suggestions to ignore logarithms in p-surprisals, which were designed to be closer to statistical practice. Indeed, from the practical point of view logarithms are important even for binary sequences: e.g., if we toss a coin (possibly biased) $N := 10^3$ times and get exactly half "heads", we will have statistically significant evidence that the tosses are not IID, since the largest product probability Q^N of this very simple event is about 2.52%; on the other hand, the sequence of outcomes may be perfectly exchangeable. The value of the largest probability can be computed from

$$\mathbb{P}^{\mathrm{R}}(A) = \frac{N!}{(N/2)!^2}2^{-N} \le 1 = \mathbb{P}^{\mathrm{X}}(A), \tag{21}$$

where A is the event of observing $N/2$ "heads" (assuming N even).

Remark 8. The case of a fixed $K := |\mathbf{Z}|$ with variable N corresponds to the bottom left corner of Fig. 1. According to (20) (and the end of Remark 7), the slope of the graph of $\log_{10} C$ in that corner becomes infinite "under microscope" as $N \to \infty$; namely, we expect the slope to grow as $\frac{1}{2}\log_{10} N$. For $N = 10^3$ and $N = 10^6$, as used in Fig. 1, this expression gives the slopes of 1.5 and 3, respectively; more precise values given by (21) are 1.598 and 3.098, respectively.

Figure 2 complements Fig. 1 by plotting $\log_{10} C$ as function of N for a fixed K explicitly. It shows two ranges for N, $N \le 10^3$ in the left panel and $N \le 10^6$ in the right panel; the two panels look similar, and the description to follow is applicable to either. Let $\mathcal{N} \in \{10^3, 10^6\}$ be the upper limit of the range of N. The base graph shows $\log_{10} C$ in black as function of N for $K \ge \mathcal{N}$. The remaining 9 graphs consist of two pieces, black and coloured; the graphs and their labels in the legends are shown in the same order, top to bottom. Let me describe, for concreteness, the bottom one labelled as $K/\mathcal{N} = 0.1$; this description will also be applicable, *mutatis mutandis*, to the other 8 graphs. The graph corresponds to $K = 0.1\mathcal{N}$ and consists of two parts: the values for $N \le K$ are shown in black and the values for $N > K$ in olive. The behaviour of the graph changes drastically after $N = K$, which is marked by using a different colour: the black part grows exponentially fast, while the olive part grows only polynomially fast (albeit for a polynomial of a high degree, namely $\lceil (K-1)/2 \rceil \ge 50$ according to (20)).

[2] In Russian, "логарифмов не надо бояться, так же как и констант"; recorded by Shen [24, note 12] and translated by the authors of [24, arXiv version].

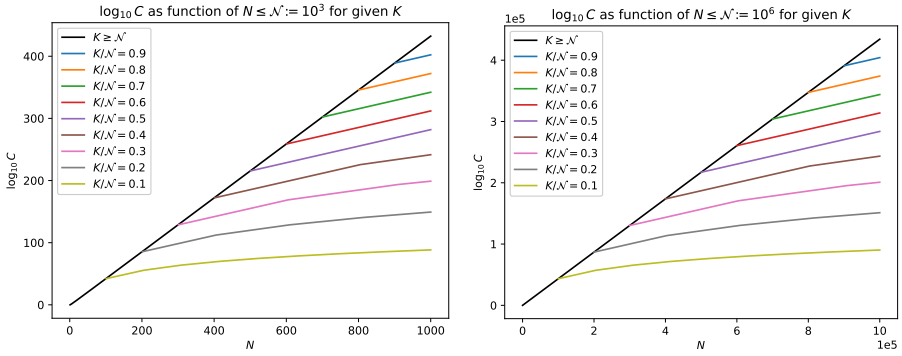

Figure 2: The graphs of $\log_{10} C$ for $\mathcal{N} = 10^3$ (left panel) and $\mathcal{N} = 10^6$ (right panel, using 10^5 as the unit for the labels on both axes), as described in text.

Let us see why the graphs in Fig. 2 look the way they do. The black graph is the diagonal of the bounding rectangle in the limit $\mathcal{N} \to \infty$. Suppose $K = \frac{k}{10}\mathcal{N}$, $k \in \{1, \dots, 9\}$, is one of the K marked in Fig. 2. If $nK < N < (n+1)K$, all n_k in a balanced split (17) are either n or $n+1$. Incrementing N by 1 to $N+1$ leads to replacing one of the $n_k = n$ by $n_k = n+1$. The relative increment in the constant C given by (16) is

$$\left(\frac{(N+1)^{N+1}}{(N+1)!}\,\frac{(n+1)!}{(n+1)^{n+1}}\right) \Big/ \left(\frac{N^N}{N!}\,\frac{n!}{n^n}\right) = \left(1+\frac{1}{N}\right)^N \left(1+\frac{1}{n}\right)^{-n} \approx \mathrm{e}\left(1+\frac{1}{n}\right)^{-n} \tag{22}$$

(the "$\approx$" is justified by $N > K$ and all the K marked in Fig. 2 being large, at least 100).

For $n = 1$, the last expression in (22) gives the slope of $\log_{10} \mathrm{e} - \log_{10} 2 \approx 0.133$ between K and $2K$. This is the slope of the full coloured lines for $K = 0.9\mathcal{N}$ (blue) to $K = 0.5\mathcal{N}$ (purple) and the slope of the first straight segment of the other coloured lines (strictly speaking, "straight" should be understood as "approximately straight" because of the "$\approx$" in (22)). The slope of the following straight segment of the coloured lines for $K = 0.4\mathcal{N}$ (brown) to $K = 0.1\mathcal{N}$ (olive) is $\log_{10} \mathrm{e} - 2\log_{10} 1.5 \approx 0.082$. The slope of the following straight segment for the bottom three coloured lines is $\log_{10} \mathrm{e} - 3\log_{10}(4/3) \approx 0.059$, etc. It is clear from (22) that the slope tends to 0 when $n \to \infty$, and we can see that the bottom line of Fig. 2 is close to being horizontal on the right. If any of the coloured lines is continued to the right beyond $\mathcal{N}$, it will consist of segments of exponentially fast growth with decreasing growth rates, which will make the overall growth rate polynomial.

5 Conclusion

Being motivated by the foundations of probability and statistics, de Finetti, Kolmogorov, and Martin-Löf considered cases where the assumptions of exchangeability and randomness are close to each other. However, there are also cases where the closeness of the two assumptions disappears, including the important case of finite sequences of a given length for a large observation space.

De Finetti's theorem has many fascinating generalizations and variations, and we can ask similar questions about those. One generalization that is especially close to the subject of this paper concerns weighted exchangeability [2], which accounts for a known covariate shift. Many more are provided by the theory of repetitive structures; see, e.g., [19], [28, Part IV], and [3, Chap. 4] (the last book, however, does not use the terminology of repetitive structures).

Acknowledgments

Thanks to the participants of the workshop "Algorithmic Statistics" (Oxford, November 28, 2025) for a useful discussion and to Ioannis Kontoyiannis for informing me about [12] and other related work.

I acknowledge the use of Microsoft Copilot in exploring proof ideas, which I reviewed carefully. I take full responsibility for this paper's mathematical statements and their proofs.

References

[1] ATLAS Collaboration. Latest results from ATLAS Higgs search. Press statement of July 4, 2012 (available on the Internet in Jan 2026).

[2] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. De Finetti's theorem and related results for infinite weighted exchangeable sequences. *Bernoulli*, 30:3004–3028, 2024.

[3] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 2000.

[4] Bruno de Finetti. La prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. An English translation of this article is included in [18] (both first and second editions).

[5] Claude Dellacherie and Paul-André Meyer. *Probabilities and Potential B: Theory of Martingales*. North-Holland, Amsterdam, 1982. Chapters V–VIII. French original: 1980; reprinted in 2008.

[6] Persi Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36:271–281, 1977.

[7] Persi Diaconis and David A. Freedman. Finite exchangeable sequences. *Annals of Probability*, 8:745–764, 1980.

[8] Lester E. Dubins and David A. Freedman. Exchangeable processes need not be mixtures of independent, identically distributed random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 48:115–132, 1979.

[9] Sander Greenland. Contribution to the discussion in [10]. *Journal of the Royal Statistical Society B*, 86:1148–1149, 2024.

[10] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing (with discussion). *Journal of the Royal Statistical Society B*, 86:1091–1171, 2024.

[11] Edwin Hewitt and Leonard J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501, 1955.

[12] Oliver Johnson, Lampros Gavalakis, and Ioannis Kontoyiannis. Relative entropy bounds for sampling with and without replacement. *Journal of Applied Probability*, 62:1578–1593, 2025.

[13] G. Jay Kerns and Gábor J. Székely. De Finetti's theorem for abstract finite exchangeable sequences. *Journal of Theoretical Probability*, 19:589–608, 2006.

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

[15] Andrei N. Kolmogorov. On tables of random numbers. *Sankhyā. Indian Journal of Statistics A*, 25:369–376, 1963.

[16] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7, 1965. Russian original: Три подхода к определению понятия "количество информации", published in Проблемы передачи информации, 1965.

[17] Andrei N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, IT-14:662–664, 1968. Russian version: К логическим основам теории информации и теории вероятностей, published in Проблемы передачи информации, 1969. The quote in Sect. 3.2 follows the translation by Alexei B. Sossinsky of this paper published as Chap. 12 "To the logical foundations of the theory of information and probability theory" (pp. 203–207) of *Selected Works of A. N. Kolmogorov*, Volume III, Information Theory and the Theory of Algorithms, edited by A. N. Shiryayev. Kluwer, Dordrecht, 1993.

[18] Henry E. Kyburg, Jr and Howard E. Smokler, editors. *Studies in Subjective Probability*. Krieger, Huntington, NY, second edition, 1980. First edition: Wiley, New York, 1964.

[19] Steffen L. Lauritzen. *Extremal Families and Systems of Sufficient Statistics*. Springer, New York, 1988.

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

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

[22] Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1–2):1–390, 2025.

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

[24] Alexey Semenov, Alexander Shen, and Nikolay Vereshchagin. Kolmogorov's last discovery? (Kolmogorov and algorithmic statistics). *Theory of Probability and Its Applications*, 68:582–606, 2024. Also published as arXiv:2303.13185 [math.LO], October 2023.

[25] Glenn Shafer. Comment on "A logic of probability, with application to the foundations of statistics" by V. Vovk. *Journal of the Royal Statistical Society B*, 55:348, 1993.

[26] Vladimir Vovk. On the concept of the Bernoulli property. *Russian Mathematical Surveys*, 41:247–248, 1986. Another English translation with proofs: arXiv:1612.08859 (math.ST).

[27] Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, San Francisco, 1999. Morgan Kaufmann.

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

A Derivation of $L(1) = 1$

In this appendix we check the statement made in Remark 7, which is equivalent to $L(1) = 1$. We can simplify the expression for $L'(x)$ in (19) as

$$L'(x) = \ln(n!) + n(n+1)\ln\frac{n+1}{n} - n\ln(n+1).$$

Integrating L' from 0 to 1 gives

$$L(1) = \sum_{n=1}^{\infty}\left(\frac{\ln(n!)}{n(n+1)} + \ln\frac{n+1}{n} - \frac{\ln(n+1)}{n+1}\right).$$

We are required to show that the partial sums

$$S_N := \sum_{n=1}^{N} \frac{\ln(n!)}{n(n+1)} + \sum_{n=1}^{N} \ln \frac{n+1}{n} - \sum_{n=1}^{N} \frac{\ln(n+1)}{n+1} \tag{23}$$

converge to 1 as $N \to \infty$. By telescoping, we can transform the first addend in (23) as

$$\begin{aligned}\sum_{n=1}^{N} \frac{\ln(n!)}{n(n+1)} = \sum_{n=1}^{N} \sum_{k=1}^{n} \frac{\ln k}{n(n+1)} = \sum_{k=1}^{N} \sum_{n=k}^{N} \frac{\ln k}{n(n+1)} \\ = \sum_{k=1}^{N} \ln k \left(\frac{1}{k} - \frac{1}{N+1} \right) = \sum_{k=1}^{N} \frac{\ln k}{k} - \frac{1}{N+1} \sum_{k=1}^{N} \ln k\end{aligned}$$

(this uses $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$ in the third equality) and the second addend as

$$\sum_{n=1}^{N} \ln \frac{n+1}{n} = \ln(N+1).$$

Plugging this into (23) gives

$$\begin{aligned}S_N &:= \sum_{k=1}^{N} \frac{\ln k}{k} - \frac{1}{N+1} \sum_{k=1}^{N} \ln k + \ln(N+1) - \sum_{n=1}^{N} \frac{\ln(n+1)}{n+1} \\ &= -\frac{\ln(N+1)}{N+1} - \frac{\ln(N!)}{N+1} + \ln(N+1) \sim 1,\end{aligned}$$

where the "=" is obtained by combining the first and last addends in the preceding expression and the definition of $N!$, and the "$\sim$" is obtained from Stirling's formula in the crude form $\ln(N!) = N \ln N - N + O(\ln N)$.