Arbor

Kolmogorov and Probability Theory

David Nualart

Arbor CLXXVIII, 704 (Agosto 2004), 607-619 pp.

1 Foundations of Probability

There is no doubt that the most famous and influential work by Andrei Nikolaevitch Kolmogorov (1903-1987) is a monograph of around 60 pages published in 1933 by Springer in a collection of texts devoted to the modern theory of probability [16]. This monograph changed the character of the calculus of probability, moving it from a collection of calculations into a mathematical theory.

The basic elements of **Kolmogorov´s** formulation are the notion of probability space associated with a given random experience and the notion of random variable. These notions were formulated in the context of measure theory. More precisely, a probability space is a measure space with total mass equal to one and a random variable is a real-valued measurable function:

(i) A *probability space* $(\Omega, \mathcal{F}, P)$ **is a triple formed** by a set Ω, a σ-field of subsets of Ω, denoted by $\mathcal{F}$, and a measure P on the measurable space $(\Omega, \mathcal{F})$ such that $P(\Omega) = 1$. The set Ω has no structure and represents the set of all possible outcomes of the random experience. The elements of $\mathcal{F}$ are the events related with the experience, and for any event $A \in \mathcal{F}$, $P(A)$ is a number in $[0, 1]$ which represents its probability. The set Ω was also called by Kolmogorov the space of elementary events.

(ii) A *random variable* is a mapping $X : \Omega \to \mathbb{R}$ such that for any real number a, the set $\{\omega \in \Omega : X(\omega) < a\}$ is an event, that is, it belongs

to the σ-field $\mathcal{F}$. A random variable X induces a probability in the Borel σ-field of the real line denoted by P_X and given by

$$P_X(B) = P(X^{-1}(B))$$

for any Borel subet B of the real line. The probability P_X is called the *law or distribution* of the random variable X.

The Russian translation of Kolmogorov's monograph appeared in 1936, and the first English version was published in 1950: *Foundations of the Theory of Probability.* The delay in the English translation shows that the formulation proposed by Kolmogorov was not immediately accepted. This fact may seem surprising in view of the noncontroversial nature of Kolmogorov's approach and its great influence in the development of probability theory. In addition, Kolmogorov's axioms were more practical and useful than other formalizations of probability, like the theory of "collectives" introduced by von Mises in 1919 (see [19]). Von Mises attempted to formalize the typical properties of a sequence obtained by sampling a sequence of independent random variables with a common distribution. Although this is an appealing conceptual problem, this construction is too awkward and limited to provide a basis for modern probability theory. So, in spite of some objections on Kolmogorov's approach that appeared at the beginning, it was definitely adopted by the young generation of probabilists of the fifties, and measure theory was proved to be a fruitful and powerful tool to describe the probability of events related to a random experience.

One of the main features of Kolmogorov's formulation is to provide a precise probability space for each random experience, and this permitted to eliminate the ambiguity caused by the multiple paradoxes in the calculus of probability like those of Bertrand and Borel. As an illustration of the power of his formalism, Kolmogorov solves in his monograph *Borel's paradox* about a random point on the sphere. Such a point is specified by its longitude $\theta \in [0, \pi)$, so that θ determines a complete meridian circle, and its latitude $\phi \in (-\pi, \pi]$. If we condition by the information that the point lies on a concrete meridian (θ is fixed), its latitude is not uniform over $(-\pi, \pi]$, and it has the conditional density $\frac{1}{4}|\cos\phi|$, whereas if we assume that the point lies on the equator (ϕ is 0 or π), its longitude has a uniform distribution on $(0, \pi)$. Since great circles are indistinguishable both statements are in apparent contradiction, and this shows the inadmissibility of conditioning with respect to an isolated event or probability zero (see Billingsley [2]).

Kolmogorov's construction of conditional probabilities using the techniques of measure theory avoids these contradictions.

The strength of Kolmogorov's monograph lies on the use of a totally abstract framework, in particular, the set or possible outcomes Ω is not equipped with any topological structure. This does not imply that in some particular problems, like the convergence or probability laws, it is convenient to work on better spaces through the use of image measures. In that sense, Kolmogorov picks up the heritage of Borel who was the pioneer in the use of measure theory and Lebesgue integral in dealing with probability problems.

We will now describe some of the main contributions of Kolmogorov's monograph:

1.1 Construction of a probability on an infinite product of spaces

At the beginning of the thirties, a great number of works of the Russian probability school were oriented to the study of stochastic processes in continuous time. In this context, the following theorem proved by Kolmogorov provides a fundamental ingredient for the formalization of stochastic processes. We recall that a stochastic process is a continuous family of random variables $\{X(t), t \geq 0\}$.

Theorem 1 *Consider a family of probability measures $p_{t_1,\ldots,t_n}$ on $\mathbb{R}^n$, $n \geq 1$, $0 \leq t_1 < \cdots < t_n$, which satisfies the following compatibility condition: Given two sets of parameters $\{t_1, \ldots, t_n\} \subset \{s_1, \ldots, s_m\}$, $p_{t_1,\ldots,t_n}$ is the corresponding marginal of $p_{s_1,\ldots,s_m}$. Then, there is a unique probability P on $\Omega = \mathbb{R}^{[0,+\infty)}$ such that for each $n \geq 1$, and each $0 \leq t_1 < \cdots < t_n$, $p_{t_1,\ldots,t_n}$ concides with the image of P by the natural projection:*

$$\begin{aligned} \mathbb{R}^{[0,+\infty)} &\longrightarrow \mathbb{R}^n \\ x &\longrightarrow (x_{t_1}, \ldots, x_{t_n}). \end{aligned}$$

In the above theorem an element $x \in \mathbb{R}^{[0,+\infty)}$ is a function $x : [0, +\infty) \to \mathbb{R}$ that can be interpreted as a trajectory of a given stochastic process. As a consequence, the law of stochastic process $\{X(t), t \geq 0\}$ is determined by the marginal laws

$$P_{(X(t_1),\ldots,X(t_n))} = P \circ (X(t_1), \ldots, X(t_n))^{-1},$$

that can be chosen in an arbitrary way.

As precedents of this theorem we can first mention the construction of a probability on $\mathbb{R}^{\mathbb{N}}$ as the product of a countable family of probabilities on the real line, done by Daniell in 1919, corresponding to the probability context of independent trials, not necessarily with a common distribution. On the other hand, using the techniques developed by Daniell, Wiener [22] constructed the probability law of the Brownian motion on the space of continuous functions.

1.2 Contruction of conditional probabilities

Applying the techniques of measure theory, Kolmogorov constructed the conditional probability by a random variable X. We present here this construction using the modern notation of conditional probabilities.

We recall first the classical definition of the conditional probability of an event C by an event D such that $P(D) > 0$:

$$P(C|D) = \frac{P(C \cap D)}{P(D)}. \tag{1}$$

Suppose we are given an event $A \in \mathcal{F}$ and a random variable X. We would like to compute the conditional probability $P(A|X = x)$. If the random variable X is continuous, the event $\{X = x\}$ has probability zero and the conditional probability $P(A|X = x)$ is not well-defined by formula (1). The conditional probability $P(A|X = x)$ should be a function $f_A(x)$ defined on the range of the random variable X, such that for any Borel subset $B \subset \mathbb{R}$ with $P(X \in B) > 0$

$$P(A|X \in B) = \int_\Omega f_A(X) dP(\cdot|X \in B). \tag{2}$$

The solution to this problem is obtained by choosing f_A as the Radon-Nikodym density of the measure $B \to P(A \cap X^{-1}(B))$, with respect to P_X, that is

$$f_A(x) = \frac{dP(A \cap X^{-1}(\cdot))}{dP_X}(x).$$

In fact, for any Borel set $B \subset \mathbb{R}$

$$P(A \cap X^{-1}(B)) = \int_B f_A(x) dP_X(x) = \int_\Omega \mathbf{1}_{\{X \in B\}} f_A(X) dP \ ,$$

and, hence, dividing both members of this equality by $P(X \in B)$ we obtain (2). In particular, if $A = X^{-1}([a,b])$, then $f_A = \mathbf{1}_{[a,b]}$. Using the modern language of conditional expectation we can write

$$f_A(X) = E(\mathbf{1}_A|X).$$

The use of measure theory allowed Kolmogorov to formulate in a rigorous way the conditioning by events of probability zero like $\{X = x\}$. From the above definition, Kolmogorov proved all classical properties of conditional probabilities.

1.3 The 0-1 law

Kolmogorov's precise definitions made it possible for him to prove the so-called 0-1 law. Consider a sequence $\{X_n, n \geq 1\}$ of independent random variables. For each $n \geq 1$ we denote by $\mathcal{G}_n = \sigma(X_n, X_{n+1}, \ldots)$ the σ-field generated by the random variables $\{X_k, k \geq n\}$. The sequence of σ-fields $\mathcal{G}_n$ is decreasing and its intersection is called the *asymptotic σ-field:*

$$\mathcal{G} = \bigcap_{n \geq 1} \mathcal{G}_n.$$

Theorem 2 (0-1 Law) *Any event in the asymptotic σ-field $\mathcal{G}$ has probability zero or one.*

A simple proof of this result, using modern notation, is as follows. Let $A \in \mathcal{G}$, and suppose that $P(A) > 0$. For any $n \geq 1$, the σ-fields $\sigma(X_1, \ldots, X_n)$ and $\mathcal{G}_{n+1}$ are independent. As a consequence, if $B \in \sigma(X_1, \ldots, X_n)$, the events A and B are independent because $A \in \mathcal{G} \subset \mathcal{G}_{n+1}$. Hence,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = P(B).$$

Therefore, the probabilities $P(\cdot|A)$ and P coincide on the σ-field $\sigma(X_1, \ldots, X_n)$ for each $n \geq 1$, and this implies that they coincide on the σ-field generated by all the random variables X_n. So, $P(A|A) = P(A)$, which implies $P(A)^2 = P(A)$ and $P(A) = 1$.

Theorem 4 (Three Series Theorem) *Let $\{X_n, n \geq 1\}$ be a sequence of independent random variables. For any constant $K > 0$ define the truncated sequence*

$$Y_n = X_n \mathbf{1}_{\{|X_n| \leq K\}}.$$

Then, the series $\sum_{n\geq 1} X_n$ converges almost surely if and only if the following three series are convergent:

$$\sum_{n\geq 1} P\big(|X_n| > K\big) = \sum_{n\geq 1} P\big(X_n \neq Y_n\big),$$

$$\sum_{n\geq 1} E(Y_n),$$

$$\sum_{n\geq 1} \mathrm{Var}(Y_n).$$

The convergence of the series $\sum_{n\geq 1} P(X_n \neq Y_n)$, implies, by the Borel-Cantelli lemma, that

$$P(\limsup \{X_n \neq Y_n\}) = 0,$$

that is, $P(\liminf \{X_n = Y_n\}) = 1$, which means that the sequences $\{X_n\}$ and $\{Y_n\}$ coincide except for a finite number of terms. So, they are equivalent and the series $\sum_{n\geq 1} X_n$ converges if and only if $\sum_{n\geq 1} Y_n$ does.

As a consequence of the above theorem, the almost sure convergence is equivalent to the convergence in probability for a series of independent random variables.

The *Law of Large Numbers* says that the arithmetic mean of a sequence of independent and indentically distributed random variables converges to the expectation. It is a fundamental result in probability theory. In the particular case of Bernoulli random variables, that is, indicator functions of events, the Law of Large Numbers asserts the convergence of the relative frequence to the probability of an event in the case of a series of independent repetitions of the experience.

The first contribution by Kolmogorov to the law of large numbers is the following result published in 1930 ([14]):

Theorem 5 *Let $\{X_n, n \geq 1\}$ be a sequence of centered independent random variables. Set $S_n = X_1 + \cdots + X_n$. then,*

$$\sum_{n\geq 1} \frac{E(X_n^2)}{n^2} < \infty \Longrightarrow \frac{S_n}{n} \longrightarrow 0 \quad \textit{almost surely.}$$

The condition on the convergence of the variances is optimal. In the case of independent and identically distributed random variables, Kolmogorov provides in [16] a definitive answer to the problem of finding necessary and sufficient conditions for the validity of the strong law of large numbers.

Theorem 6 (Strong Law of Large Numbers) *Let $\{X_n, n \geq 1\}$ be a sequence of independent random variables with the same distribution. Then,*

$$E(|X_1|) < \infty \Longrightarrow \frac{S_n}{n} \longrightarrow E(X_1) \quad \textit{almost surely.}$$

$$E(|X_1|) = \infty \Longrightarrow \limsup_{n\to\infty} \frac{|S_n|}{n} = +\infty \quad \textit{almost surely.}$$

Suppose that $\{X_n, n \geq 1\}$ is a sequence of centered, independent random variables with the same distribution. Set $S_n = X_1 + \cdots + X_n$, for each $n \geq 1$. The Strong Law of Large Numbers says that

$$\frac{S_n}{n} \longrightarrow 0,$$

almost surely. On the other hand, if $E(X_1^2) < \infty$, the Central Limit Theorem asserts that $\frac{S_n}{\sqrt{n}}$ converges in distribution to the normal law $N(0, \sigma^2)$, where $\sigma^2 = E(X_1^2)$, that is, for any real numbers $a \leq b$

$$P\left(a \leq \frac{S_n}{\sqrt{n}} \leq b\right) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx.$$

Taking into account these results, one may wonder about the asymptotic behaviour of S_n as n tends to infinity. The Law of Iterated Logarithm, established by Khintchine in 1924 ([9]), precises this behavior:

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log n}} = \sigma^2$$

almost surely. In 1929 Kolmogorov proved in [12] the following version of the law of iterated logarithm for non identically distributed random variables.

Theorem 7 *Let $\{X_n, n \geq 1\}$ be a sequence of independent random variables with zero mean and finite variance. Set $S_n = X_1 + \cdots + X_n$, for each $n \geq 1$. Then,*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2B_n \log\log B_n}} = 1$$

almost surely, where

$$B_n = \sum_{k=1}^{n} E(X_k^2) \uparrow \infty$$

and

$$|X_n| \leq M_n = o\left(\sqrt{B_n \log\log B_n}\right).$$

The proof of this result given by Kolmogorov makes possible its extension to unbounded random variables through a truncation argument. This proof can be considered as modern in the sense that it introduces new techniques like the large deviations, which have become fundamental.

3 Stochastic processes

In 1906-1907, Markov [17, 18] discovered that limit theorems for independent random variables could be extended to variables "connected in a chain". About the same time, Einstein [7] published his work on Brownian motion. In this context, the celebrated work by Kolmogorov ([15]) sinthesized these researches and was the starting point of the theory of Markov processes. The modern definition of a Markov process is as follows:

Definition 8 *A stochastic process* $\{X_t, t \geq 0\}$ *with values in a state space* E *is called a Markov process if for any* $s < t$ *and any measurable set of states* $A \subset E$ *it holds that*

$$P\left(X_t \in A | X_r, 0 \leq r \leq s\right) = P(X_t \in A | X_s).$$

The heuristic meaning of this definition is the independence of the future and past values of the process if we know its present value. Kolmogorov called these processes "stochastically determined processes". The name of Markov processes was suggested by Khintchine in 1934.

We can write

$$P(X_t \in A | X_s) = P(s, X_s, t, A),$$

and the function $P(s, x, t, A)$ describes the probability that the process is in A at time t, conditioned by the information that at time s is in x. Thus, $A \longrightarrow P(s, x, t, A)$ is a family of probabilities parameterized by s, x and t

called the *transition probabilities* of the process. They satisfy the so-called Chapman-Kolmogorov equation:

$$P(s,x,t,A) = \int_E P(s,x,u,dy)P(u,y,t,A), \tag{4}$$

for any $s < u < t$, and they allow to describe all probabilistic properties of the process. Chapman has mentioned this equation in the work [4] on Brownian motion in 1928.

Kolmogorov's approach to Markov process developed in [15] is purely analytic and the main goal is to find regularity conditions on the transition probabilities $P(s,x,t,dy)$ in order to handle the Chapman-Kolmogorov equation (4). The central ideal of Kolmogorov's paper is the introduction of local characteristics at time t and the construction of transition probabilities by solving certain differential equations involving these characteristics. In the case of real-valued processes (that is, $E = \mathbb{R}$), Kolmogorov considers the class of transition functions for which the following limits exist

$$\begin{aligned} A(t,x) &= \lim_{\delta \downarrow 0} \frac{1}{\delta} \int_{\mathbb{R}} (y-x)P(t,x,t+\delta,dy), \\ B(t,x) &= \lim_{\delta \downarrow 0} \frac{1}{2\delta} \int_{\mathbb{R}} (y-x)^2 P(t,x,t+\delta,dy). \end{aligned}$$

Feller suggested the names *drift* and *diffusion coefficients* for these limits. A property on the third moments is also needed to exclude the possibility of jumps. Assuming, in addition, that the density function of the measure $P(s,x,t,\cdot)$, denoted by

$$f(s,x,t,y) = \frac{P(s,x,t,dy)}{dy},$$

is sufficiently smooth, Kolmogorov proved that it satisfies the *forward differential equation*:

$$\frac{\partial f}{\partial t} = -\frac{\partial [A(t,y)f]}{\partial y} - \frac{\partial^2 [B^2(t,y)f]}{\partial y^2} \tag{5}$$

and the *backward differential equation*:

$$\frac{\partial f}{\partial s} = -A(s,x)\frac{\partial f}{\partial x} - B^2(s,x)\frac{\partial^2 f}{\partial x^2}.$$

Equation (5) arises if the study of the time evolution of the probability distribution of the process and a special form of this equation appeared earlier in papers of Fokker [8] and Planck [20]. Kolmogorov called Equation (5) the Fokker-Planck equation since 1934. When the coefficients depend only on time (processes homogeneous in space), these equations appeared first in 1900 in a paper of Bachelier [1].

The construction of transition functions from the drift and diffusion coefficients motivated the works on fundamental solutions to parabolic partial differential equations and were the starting point on the fruitful relationship between Markov processes and parabolic equations.

Although his point of view on the theory of stochastic processes was mainly analytical, Kolmogorov also developed a certain number of tools for the study of the properties of the paths of stochastic processes. Among these tools, the most famous and most used is the criterion that guarantees the continuity of the trajectories of a given stochastic process from conditions on the moments of its increments. This criterion was proved by Kolmogorov in 1934 and presented in the Seminar of Moscow University. However, Kolmogorov never published this result, and it was Slutsky who stated and give the first proof in [21] in 1934, attributing it to Kolmogorov.

Definition 9 *We say that two stochastic processes* $\{X_t, 0 \leq t \leq 1\}$ *and* $\{Y_t, 0 \leq t \leq 1\}$ *are equivalent (or that* X *is a version of* Y*) if for any* t, $P(X_t \neq Y_t) = 0$.

Two equivalent processes may have different trajectories.

Theorem 10 (Kolmogorov continuity criterion) *Consider a stochastic process* $\{X_t, 0 \leq t \leq 1\}$. *Suppose that there exists constants* $p, c, \varepsilon > 0$ *such that*

$$E\left(|X_t - X_s|^p\right) \leq c|t-s|^{1+\varepsilon},$$

for all $s, t \in [0,1]$. *Then, there exists a version of the process* X *with continuous trajectories.*

One can also show under the same hypotheses that there exists a version of the process X with Hölder continuous trajectories of order $\alpha < \frac{1+\varepsilon}{p}$, that is,

$$|X_t - X_s| \leq G|t-s|^{\alpha}$$

for all $s, t \in [0,1]$, and for some random variable G.

As an example of the application of this theorem, consider the case of the Brownian motion $\{B_t, 0 \leq t \leq 1\}$, defined as a Gaussian process (its finite dimensional distributions are Gaussian) with zero mean and covariance function $E(B_t B_s) = \min(s,t)$. This process has stationary and independent increments and the law of an increment $B_t - B_s$ is $N(0, t-s)$. As a consequence, for any integer $k \geq 1$

$$E(|B_t - B_s|^{2k}) = \frac{(2k)!}{k! 2^k} |t-s|^k,$$

and there is a version of the Brownian motion with Hölder continuous trajectories of order α, for any $\alpha < \frac{1}{2}$.

4 Conclusions

A) Kolmogorov may be considered as the founder of probability theory. The monograph by Kolmogorov published in 1933 transformed the calculus of probability into a mathematical discipline. Some authors compare this role of Kolmogorov with the role played by Euclides in geometry.

B) The results on limit theorems for sequences and series of independent random variables established by Kolmogorov were definitive and constitute a basic core of results on any text course in probability.

C) Kolmogorov ideas influenced decisively almost all the work on Markov processes and make possible the posterior development of stochastic analysis.

References

[1] L. Bachelier. Théorie de la spéculation. *Ann. Sci. École Norm. Sup.* **17**, 21-86 (1900).

[2] P. Billingsley. *Probability and Measure,* John Wiley 1979.

[3] L. Chaumont, L. Mazliak and M. Yor: A. N. Kolmogorov. *Quelques aspects de l'oeuvre probabiliste.* In "L'héritage de Kolmogorov en mathématiques". Ed. Berlin, collection Échelles, 2003.

[4] S. Chapman. On the Brownian displacement and thermal diffusion of grains suspended in a non-uniform fluid. *Proc. Roy. Soc. London, Ser. A* **119**, 34-54 (1928).

[5] J. L. Doob. Kolmogorov's early work on convergence theory and foundations. *Ann. Probab.* **17**, 815-821 (1989).

[6] E. B. Dynkin. Kolmogorov and the theory of Markov processes. *Ann. Probab.* **17**, 822-832 (1989).

[7] A. Einstein. Zur Theorie des Brownschen Bewegung. *Ann. Physik* **19**, 371-381 (1906).

[8] A. D. Fokker. Die mittlere Energie rotierende elektrischer Dipole im Strahlungsfeld. *Ann. Physik* **43** 810-820 (1914).

[9] A. Y. Khinchin. Über einen Satz der Wahrscheinlichkeitrechnung. *Fund. Mat.* **6**, 9-20 (1924).

[10] A. Y. Khinchin et A. N. Kolmogorov. Über Konvergenz von Reihen, deren Glieder durch den Zufall bestimmt werden. *Mat. Sb.* **32**, 668-677 (1925).

[11] A. N. Kolmogorov. Über dir Summen durch den Zufall bestimmter unabhängiger Grössen. *Math. Ann.* **99**, 309-319 (1928).

[12] A. N. Kolmogorov. Über das Gesetz des iterierten Logarithmus. *Math. Ann.* **101**, 126-135 (1929).

[13] A. N. Kolmogorov. Bermerkungen zu meiner Arbeit "Über die Summen zufälliger Grössen". *Math. Ann.* **100**, 4844-488 (1930).

[14] A. N. Kolmogorov. Sur la loi forte des grands nombres. *CRAS Paris* **191**, 910-912 (1930).

[15] A. N. Kolmogorov. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math.Ann.* **104**, 415-458 (1931).

[16] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer 1933.

[17] A. A. Markov. Extension of the law of large numbers to independent events. *Bull. Soc. Phys. Math. Kazan* **15**, 135-156 (1906). (In Russian).

[18] A. A. Markov. Extension of limit theorems of probability theory to a sum of variables connected in a chain. *Zap. Akad. Nauk. Fiz.-Mat. Otdel., Ser. VIII* **22**. Presented Dec. 5 1907.

[19] R. von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Z.* **5**, 52-99 (1919).

[20] M. Plank. Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. *Sitzungsber. Preuss. Akad. Wiss. Phys.-Math. Kl.* 324-341 (1917).

[21] E. B. Slutsky. Qualche proposizione relativa alla teoria delle funzioni aleatorie. *Giornale dell'Instituto Italiano dei Attuari* **8**, 183-199 (1937).

[22] N. Wiener. Differential space. *Jour. Math. and Phys.* **58**, 131-174 (1923).