The Annals of Probability
1989, Vol. 17, No. 3, 815–821

KOLMOGOROV'S EARLY WORK ON CONVERGENCE THEORY AND FOUNDATIONS

BY J. L. DOOB

University of Illinois, Urbana-Champaign

1. Probability before Kolmogorov. When Kolmogorov was starting his mathematical career, nonmathematical probability was, as it still is, the study of various not very precisely defined real contexts. Some of these contexts gave rise to mathematical problems, in combinatorics for example, but it was not clear what an overall mathematical probability context would be or indeed whether one was possible. Poincaré had written in 1912 [13] "On ne peut donner une définition satisfaisante de la probabilité." It was typical of the writing of that time, and in fact of considerably more recent times, that the reader could not be certain whether the writer was thinking of probability as a nonmathematical or a mathematical subject in his statement. von Mises in 1919 [15] was clearer in what he deplored but was just as pessimistic, although more professorily ponderous: "In der Tat kann man den gegenwärtigen Zustand kaum anders als dahin kennzeichnen, dass die Wahrscheinlichkeitsrechnung heute ein mathematische Disciplin nicht ist." von Mises attempted to create his desired mathematical discipline but his theory of "collectives" was a confused although suggestive mixture of mathematical and nonmathematical contexts. In view of Kolmogorov's high opinion of von Mises a few explanatory remarks are appropriate here. Consider a sequence obtained by sampling a sequence of independent trials with a common distribution. (Note that sampling an infinite sequence is an unrealistic element of this analysis.) There are typical properties associated with such a sequence, as indicated for example by the law of large numbers. von Mises attempted to construct a basis for mathematical as well as nonmathematical probability by a formalization of these typical properties, by constructing an individual sequence with enough of these properties to be a model for a trial sequence. His original definition of such an individual sequence (a "collective") [15], was insightful of what seems to happen in sampling, but was either vacuous or meaningless when applied to an individual sequence, depending on the reader'sinterpretation of von Mises' words. His later definition [16] had too few properties to be useful. In any event such a construction, even if successful, would obviously be too awkward and too limited to be considered seriously as a useful basis for the extraordinary scope of modern probabilistic mathematical analysis. On the other hand, the formalization of such a sequence is an appealing conceptual problem which Kolmogorov discussed on several occasions, most recently in 1983 [K462].[1]

Received August 1988.

[1]Reference citations preceded by K refer to the list of Kolmogorov's publications on pages 945–964.

Although it was not clear when Kolmogorov was starting his research what the mathematical context of probabilistic analysis was, the existing fund of technical mathematical results was by no means negligible. For example, versions of the law of large numbers and sophisticated versions of the central limit theorem had been proved. Although most theoretical investigations were concerned only with repeated independent trials, that is, with sequences of independent random variables with a common distribution, Markov defined and discussed what are now called Markov chains in 1906 [12]. Borel, in an influential 1909 paper [3], called attention to almost sure properties of infinite sequences of trials and proved his half of the Borel–Cantelli theorem; Cantelli proved the second half in 1917 [4]. In Borel's 1909 paper he stated the strong law of large numbers for symmetric Bernoulli trials and linked it to a property of dyadic expansions of numbers in the interval $(0, 1)$, but his proof was hardly acceptable in that it was based on evaluations using the central limit theorem for Bernoulli trials with the remarkable simplifying addition that the approximation error was taken to be zero! In 1910, Faber [7] gave what seems to be the first valid proof of this case of the strong law of large numbers. He obtained the result as a property of dyadic expansions and refers to Borel's paper for the connection with probability. Apparently Hausdorff's 1914 set theory book [9] contained the first rigorous simple proof of this law and was the first of papers by several authors culminating in versions of the iterated logarithm law, of which Khintchine's in 1924 [11] was the first. Hausdorff explicitly identified probability with measure in his context, again that of dyadic expansions of numbers in $(0, 1)$. Writers in this period were sometimes chary to identify mathematical probability with measure because there were certain ideas commonly held in nonmathematical probability discussions, such as *probability* 0 *means impossibility*, that are incompatible with the measure approach.

Bachelier, in papers from 1900 [1] on, derived properties of the Brownian motion process from asymptotic Bernoulli trial properties. His Brownian motion process was necessarily not precisely defined, but his application of the André reflection principle becomes valid for the Brownian motion process as an application of the strong Markov property. His valuable results were repeatedly rediscovered by later researchers. Wiener, in a 1923 paper [17], applied the Daniell integral to give a rigorous treatment of the Brownian motion process, now accordingly sometimes called the Wiener process, but this work, like his pioneering work in potential theory, went unnoticed and unused for years.

2. Kolmogorov's classical work on sums of independent random variables. After a treatment of convergence of infinite series of discretely distributed independent random variables in his part of a 1925 joint paper [K10] with Khintchine, Kolmogorov, in a 1928 paper [K18, K23], dropped the hypothesis of discretely distributed summands and proved that an infinite series of independent mean 0 random variables converges almost surely if the series of summand variances converges. Furthermore he proved his famous Three Series Theorem, giving necessary and sufficient conditions for the almost sure convergence of infinite series of independent random variables. A fundamental tool was

his inequality, now called Kolmogorov's inequality, for maxima of partial sums of mean 0 independent random variables. This inequality generalizes Chebyshev's inequality, was later extended by Bernstein [2] to what are now called martingales and is now classified as an application of a submartingale inequality. Although the context of the inequality has widened, Kolmogorov's proof is trivially adaptable to the newer context. In [K18, K23] Kolmogorov defined equivalence of sequences of random variables: Two sequences $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ of random variables are defined as equivalent if $\Sigma_n P[x_n \neq y_n] < \infty$. This condition implies that, for almost every sample sequence, $x_n = y_n$ for sufficiently large n. Using the idea of replacing a sequence of random variables by an equivalent sequence of suitably bounded random variables, Kolmogorov in [K18, K23] proved his Three Series Theorem, which reduces a series convergence problem to one involving mean 0 and bounded summands, and reduced various laws of large numbers to corresponding laws formulated in terms of L^1 and L^2 limits for equivalent random variable sequences. As an application he proved that if σ_n is the average of the first n random variables of a sequence of independent random variables with a common distribution function F, then there is a sequence c_n of constants for which $\sigma_n - c_n \to 0$ in probability if and only if

$$(*) \qquad \lim_{n\to\infty} nP[|\sigma_1| > n] = 0.$$

He noted later, in [K40], that the constants can be chosen equal to each other, $c_1 = c_2 = \cdots = c$ if and only if, besides $(*)$, the symmetric integral

$$\lim_{b\to\infty} \int_{-b}^{b} a\, dF(a)$$

exists, and then c can be chosen as the value of this integral.

In 1929 [K21] Kolmogorov proved a version of the iterated logarithm law for sums of independent mean 0 bounded random variables. In this version, if B_n is the variance of the nth partial sum s_n and if m_n is a bound for the absolute value of the nth summand, then the conditions $B_n \to \infty$ and $m_n = o(B_n/\ln\ln B_n)^{1/2}$ imply that

$$\limsup_{n\to\infty} s_n(2B_n \ln\ln B_n)^{-1/2} = 1$$

almost surely. The first of the two conditions implies the second if the summands are uniformly bounded. This theorem makes it possible to apply random variable sequence equivalence to prove delicate iterated logarithm results for unbounded summands, by truncation. The best previous result was Khintchine's proof of the theorem [11] for Bernoulli trials.

In 1930 [K24] Kolmogorov proved one of his best known theorems, that if $x_1, x_2, \ldots$ is a sequence of independent mean 0 random variables with variances $b_1, b_2, \ldots$ and if $\Sigma_n b_n n^{-2} < \infty$, then $(x_1 + \cdots + x_n)/n \to 0$ almost surely and the variance condition cannot be weakened.

3. Kolmogorov's 1933 monograph [K40]. This influential monograph transformed the character of the calculus of probabilities, moving it into mathematics from its previous state as a collection of calculations inspired by a vague nonmathematical context, a context thought to justify the use of half-defined pseudomathematical concepts.

When Khintchine and Kolmogorov studied the convergence of infinite series of independent random variables in their 1925 paper [K10], they did not mention explicitly their hypothesis of summand independence, a hypothesis which probably seemed so natural at the time that the explicit statement was lost in the writing. They reduced their problem to one of summation of certain Lebesgue measurable functions on the unit interval $(0, 1)$ under Lebesgue measure on the interval. More precisely, using their hypothesis that their random variable summands were discretely distributed, they were able to construct, corresponding to each random variable sequence, a sequence of Lebesgue measurable functions on the unit interval with the same joint distributions in terms of Lebesgue measure as the given summand random variable joint probability distributions. Since almost sure convergence is defined in terms of these joint distributions, in the same way in the two contexts, they could solve their problem in the standard context of Lebesgue measure theory. Thus they solved a well defined mathematical problem, even though their statement of the original problem was not in properly defined mathematical terminology. When Kolmogorov studied the same convergence problem (without the discrete distribution hypothesis) in his 1928 paper [K18, K23], the independence hypothesis was in the title of the paper. Expectations were integrals of random variables and expectations with respect to an event B were integrals over the set B. But in this paper there was no explanation of the identification of events with sets of a measure space or of the random variables with measurable functions on the space. The omission was customary at the time, as was the fact that the omission was not mentioned. It was not until the 1933 monograph that the standard manipulations of probabilities and expectations were fully justified.

By the 1930's it was understood that the basic manipulations of mathematical probability were the same as those of measure theory, but the relation between the two had not been given a usable formulation. In such a formulation:

(A) The probabilistic context must be identified with a probability measure space, that is, a measure space for which the measure of the space itself is 1.
(B) Random variables and their expectations must be identified with measurable functions on the probability space and their integrals.
(C) Conditional probabilities and integrals must be defined.

Once a probability measure has been chosen, as suggested by a given nonmathematical or mathematical context, it costs nothing—except perhaps the annoyance of nonprobabilists—to call measurable functions on the space "random variables" and to call their integrals with respect to the chosen measure "expectations". Thus (B) becomes trivial. However there was a startling innovation in Kolmogorov's solution of (C) in that he unexpectedly defined conditional probabilities and expectations as random variables, whose existence he proved by an application of the Radon–Nikodym theorem.

A treatment of (A), (B) and (C) does not provide a useful basis for mathematical probability until it is shown how the treatment can be adapted to standard probability contexts. The natural mathematical space corresponding to a nonmathematical probabilistic context producing some sort of outcome is the space of all possible outcomes, called by Kolmogorov the space of elementary events. A probability measure is to be defined on some σ algebra of subsets of this space with values suggested by the context. This space is frequently a product space. For example, in a common nonmathematical probability context at each point of a parameter set T, usually identified with a set of values of time, the probability process produces a real number. In this context, the space of elementary events is the product space R^T, the class of functions from T into the reals. The mathematical problem (A) becomes that of defining a probability measure on this product space, adapted to the given context.

In 1913 Radon [14] treated measures of Borel sets in finite dimensional Euclidean space; there remained the problem of defining measures on R^T for T infinite. Kolmogorov did this in [K40]. For T countably infinite, Daniell, in 1919, had already defined product probability measures on R^T (corresponding to the probability context of independent trials, not necessarily with a common distribution [5]) and in [6] Daniell defined (in an awkward formulation) general probability measures on R^T, but his papers were not probabilistically oriented and were apparently unknown to Kolmogorov. As Kolmogorov's treatment for arbitrary T shows, the decisive case is for T countably infinite, in which case Daniell's work yields Kolmogorov's measure, but is awkwardly formulated except in the product measure case. Kolmogorov's approach is suggested by a rephrasing of the nonmathematical context described above. In this context to each point of the arbitrary set T there corresponds a real valued random variable, and every finite set of these random variables has a joint distribution prescribed by the context. A probability measure space is to be constructed on which a family $\{x(t),\ t \in T\}$ of measurable functions is to be defined with the same finite dimensional joint distributions as the given random variables. For t in T and f in R^T let $x(t)$ be the tth coordinate function on R^T, that is, $x(t)$ is the function from R^T into the reals whose value at f is $f(t)$. The coordinate functions are to be the mathematical counterparts of the nonmathematical random variables. Kolmogorov showed that there is a probability measure defined on the smallest σ algebra of subsets of R^T making every coordinate function measurable, assigning the prescribed joint distributions to the finite sets of coordinate functions. He did this without recourse to a nonmathematical context, that is, for arbitrary T and an arbitrary assignment of mutually consistent distributions of finite sets of coordinate functions, Kolmogorov defined a probability measure on his choice of σ algebra of subsets of R^T, with the assigned finite dimensional distributions.

Thus Kolmogorov necessarily had to treat measure theory on abstract spaces. Although Fréchet had noted in 1915 [8] that the concepts associated with Lebesgue measure on the line extend to measure on abstract spaces, these ideas had evidently not been fully absorbed by the time [K40] appeared. At any rate, Kolmogorov, paying his tribute to cultural lag, considered it advisable to define measurability of a function, to define the integral of a measurable function, to

prove that the limit of a convergent sequence of measurable functions is measurable and so on, in the context of an abstract measure space. It is surprising that in a discussion involving going to the limit under the expectation symbol, Kolmogorov did not appeal to the Lebesgue dominated convergence theorem.

In view of the seemingly noncontroversial nature of Kolmogorov's approach to mathematical probability, it may seem surprising that it was not universally accepted at once. Perhaps one reason for the delay was the mistaken idea that he was limiting the scope of mathematical probability by making the subject purely mathematical instead of keeping it a combination of mathematical and nonmathematical contexts. To give the flavor of the theory's reception, note that one research probabilist asserted to the writer in the 1940's that Kolmogorov's approach was not applicable to a game between two players in which the winner had to win two out of three plays, because there was not a fixed number of plays to a game! A second objection was offered by some classical analysts: They asserted that measure theory deprived probability theory of its charm, that measure theory was tedious and boring and that they doubted that it would prove fruitful as applied to probability. History has refuted these doubters, although they would have felt even more doubtful if they could have imagined the refined delicate set theoretic analysis that has invaded present day probability theory. Kolmogorov himself stated in [K40] that what distinguishes probability from the usual measure theory are such conditions as independence and "weakened analogous conditions," for example the Markov property.

Kolmogorov's new definition of conditional probabilities made it possible to define the Markov property precisely: If $x_1, x_2, \ldots$ is a sequence of random variables, his definition covers conditional probabilities relative to $(x_1, \ldots, x_n)$ for all n. In fact, however, it was many years before precise definitions became standard in the literature. Kolmogorov himself, who in his early work used the somewhat deceptive descriptive term "stochastically definite" instead of "Markov," stated the defining Markov property imprecisely even after he had stated it quite precisely, at least for discretely distributed random variables, in [K40].

There is one way in which Kolmogorov's measure conventions differ from the present ones. The present convention is that a random variable is defined as a measurable function from a specified probability measure space into a specified measurable space, and concepts like random variable distributions, random variable mutual independence, conditional expectations given a random variable and so on, depend on the specified classes of measurable sets of the range spaces of the random variables in question. Kolmogorov adopted a different convention in his definitions, essentially the following. For each function x from a probability measure space into a space, Kolmogorov made the range space of x measurable by defining its class of measurable sets as the smallest class making x measurable. His convention may lead to unexpected contradictions with present definitions. It may happen, for example (see [10]), that two real random variables, that is, two functions from a probability measure space into the reals, are independent random variables in the modern sense, in which the range space of the random variables has been made measurable by the assignment of the Borel

sets, even though these random variables are not independent under Kolmogorov's definition, which may be more demanding.

In [K40] Kolmogorov stated the best strong law of large numbers for a sequence $x_1, x_2, \ldots$ of independent random variables with a common distribution: $(x_1 + \cdots + x_n)/n$ has an almost everywhere finite limit when $n \to \infty$ if and only if $E[|x_1|] < \infty$, and if so, the limit is $E[x_1]$. Finally, Kolmogorov's precise definitions made it possible for him to prove in [K40] the following 0-1 law: If $y_1, y_2, \ldots$ is a sequence of real random variables and if f is a Baire function of this sequence, with the property that

$$P[f = 0|y_1, \ldots, y_n] = P[f = 0] \quad \text{a.s.}$$

for all n, then the right side of the equality must be either 0 or 1. (The set where $f = 0$ would now be described as a set in the σ algebra generated by $y_1, y_2, \ldots$.)

REFERENCES

[1] BACHELIER, L. (1900). Théorie de la speculation. *Ann. Sci. École Norm. Sup.* (3) **17** 21–86.

[2] BERNSTEIN, S. (1937). On some modifications of Chebychev's inequality. *Dokl. Akad. Nauk SSSR* **17** 275–278. (In Russian.)

[3] BOREL, É. (1909). Les probabilités dénombrables et leurs applications arithmétiques. *Rend. Circ. Mat. Palermo* **27** 247–271.

[4] CANTELLI, F. (1917). Sulla probabilita come limite della frequenza. *Rom. Rend. Accad. Lincei* (5) **26** 39–45.

[5] DANIELL, P. (1918 / 1919). Integrals in an infinite number of dimensions. *Ann. of Math.* (2) **20** 281–288.

[6] DANIELL, P. (1919 / 1920). Functions of limited variation in an infinite number of dimensions. *Ann. of Math.* (2) **21** 30–38.

[7] FABER, G. (1910). Über stetige Funktionen. II. *Math. Ann.* **69** 372–443.

[8] FRÉCHET, M. (1915). Sur l'intégrale d'une fonctionelle étendue à un ensemble abstrait. *Bull. Soc. Math. France* **43** 248–265.

[9] HAUSDORFF, F. (1914). *Grundzüge der Mengenlehre*. Veit, Leipzig; (1927) *Mengenlehre*, 2nd ed. de Gruyter, Berlin.

[10] JESSEN, B. (1948). On two notions of independent functions. *Colloq. Math.* **1** 214–215.

[11] KHINTCHINE, A. (1924). Über einen Satz der Wahrscheinlichkeitsrechnung. *Fund. Math.* **6** 9–20.

[12] MARKOV, A. A. (1906). Extension of the law of large numbers to dependent events. *Bull. Soc. Phys. Math. Kazan* (2) **15** 135–156. (In Russian.)

[13] POINCARÉ, H. (1912). *Calcul des probabilités*, 2nd ed. Gauthier-Villars, Paris.

[14] RADON, J. (1913). Theorie und Anwendungen der absolut additiven Mengenfunktionen. *Sitzungsber. Akad. Wiss. Wien Math.-Naturwiss. Kl.* **122** 1295–1438.

[15] VON MISES, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Math. Z.* **5** 52–99.

[16] VON MISES, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic, New York.

[17] WIENER, N. (1923). Differential space. *J. Math. Phys.* **2** 131–174.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801