Kolmogorov's Axioms and Infinite Dimensional Probability Spaces

Andrew Stewart

March 8, 2010

1 Axiomatic Probability Theory

The first Western mathematicians to study probability were Fermat and Pascal in the 17^{th}century. Bernoulli and de Moivre treated the subject more mathematically in the 18^{th}century: they postulated that the probability of an event is the ratio of the number of equally likely cases in which the event occurs to the total number of possible cases.

$$\mathbf{P}(A) = \frac{\#\text{ of cases in which } A \text{ occurs}}{\text{total }\#\text{ of cases}}$$

Their framework for probability remained, with some exceptions, as the basis for the subject until the early 20^{th}century.

The advent of measure theory at the turn of the 20^{th}century gave mathematicians an appropriate tool to formulate a more robust model of probability. Indeed, several attempts of such a formulation were made in response to Hilbert's sixth problem:

> To treat in the same manner, by means of axioms, those physical sciences in which mathematics plays an important part; in the first rank are the theory of probabilities and mechanics.

Fréchet first considered measures on an arbitrary space as functions. It was was on this idea that Kolmogorov based his probability theory in his famous text, *Grundbegrie der Wahrschein-lichkeitsrechnung*. The theory made heavy use of the work of several mathematicians: Borel, Lebesgue, Carathéodory, Radon, Nikodym. In this paper, we will briefly introduce the framework laid out by Kolmogorov and provide the outline of a proof of his extension theorem. This theorem gives a good sense of the appropriate application of measure theory to probability theory.

The axioms given in the Grundbegrie, using modern terminology, are as follows. Let Ω be a set, $\mathcal{F}$ be a set of subsets (called *events*) of Ω.

- $\mathcal{F}$ is σ-algebra of sets.

- $\Omega \in \mathcal{F}$.
- $\mathbf{P}$ is a function from $\mathcal{F}$ to $[0, \infty)$.
- $\mathbf{P}(\Omega) = 1$.
- If $A \cap B = \emptyset$ then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

This list is sufficient to describe all of classical probability theory, which only deals with finite probability spaces. To it was added a sixth axiom[1] which, given the first five, is equivalent to countable additivity (Why?):

- If $A_1 \supset A_2 \supset \ldots$ is a sequence of events in $\mathcal{F}$ such that $\mathbf{P}(\bigcap_j A_j) = 0$, it follows that $\lim_{j \to \infty} \mathbf{P}(A_j) = 0$.

When these six axioms are satisfied, we say that $(\Omega, \mathcal{F}, \mathbf{P})$ is a *probability space*; the function $\mathbf{P}$ is a *probability measure*. The simplest example of an infinite probability space is a subset of $\mathbf{R}$ of finite measure, with a properly scaled Lebesgue measure as the probability measure. This is the so-called *uniform distribution*.

In this model, random variables are simply measurable functions $X : \Omega \to \mathbf{R}$. The number

$$\mathbf{P}(\{\omega \in \Omega \mid X(\omega) \in A\}),$$

which is the probability that X lies in A, is abbreviated by

$$\mathbf{P}(X \in A).$$

Random vectors are just what one would expect them to be: functions from a probability space to $\mathbf{R}^n$.

Independence of sets is defined using the probability measure: events A and B are independent if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B);$$

σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ that are subalgebras of $\mathcal{F}$ are independent if $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$ imply that A and B are independent with respect to $\mathcal{F}$. Random variables X and Y are independent if the σ-algebra $\sigma(X)$ and $\sigma(Y)$ are independent, where

$$\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}\},\ \mathcal{B} \text{ is the Borel } \sigma\text{-algebra }.$$

It's easy to see that X and Y are independent if and only if for every real a and b, the sets $\{X \leq a\}$ and $\{Y \leq b\}$ are independent.

Integration is developed as it standardly is in measure theory. The expectation of a random variable is defined in terms of integrals:

$$\mathbf{E}X := \int X d\mathbf{P},$$

[1]Note that such an axiom is difficult to motivate based on experience from real life, for any observable random process is finite (and well described by classical probability theory). However, this axiom proved to be crucial in developing the theory.

where integration with respect to the measure $\mathbf{P}$ is defined in a completely analogous way as integration with respect to the Lebesgue measure. This definition agrees with both the classical definition in the case when $\mathbf{P}$ is discrete, and the undergraduate definition given for absolutely continuous random variables[2]. Theorems that we know about the Lebesge integral – such as Fatou's lemma, monotone convergence theorem, and dominated convergence theorem[3] – have the same proofs as for the Lebesgue measure.

Thus, probability theory is put on a firm mathematical framework.

Example. Consider a coin toss. There are two possible outcomes: heads or tails. We can model it using a probability space with two elements. Let $\Omega = \{H, T\}$ and $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$. If the coin is fair, we should have $\mathbf{P}(H) = \mathbf{P}(T) = 1/2$.

The random variable $X : H \to 1, T \to -1$ can be interpreted as the payoff of a game: the player wins one dollar if the coin is heads and loses a dollay if the coin is tails. The expectation of X is

$$\mathbf{E}(X) = 1 \cdot \frac{1}{2} + -1\frac{1}{2} = 0,$$

as it should be. Here, the probability measure has point masses at -1 and 1, so the integral ends up being a sum.

2 Kolmogorov's Extension Theorem

Classical probability theory has no difficulty modeling a finite number of coin flips. One can explicitly calculate the probability of any event. How would it model an infinite (say, countable) number? What is the probability that every tenth coin flip is a head? It is impossible to answer questions like these using the classical definition, for both the number of cases in which every tenth coin is a head and the total number of cases are infinite.

In terms of modern probability theory, we can model the flipping of n coins by the probability space $(\Omega_n, \mathcal{F}_n, \mathbf{P}_n)$ where

- $\Omega_n = \{(X_1, X_2, \ldots, X_n) \mid X_i \in \{0, 1\}, i = 1, 2, \ldots, n\}$;
- $\mathcal{F}_n = 2^{\Omega_n}$; and
- $\mathbf{P}(A) = |A|/2^n$.

It is natural to believe that the probability space required for an infinite number of coin flips would be

$$\Omega = \{(X_1, X_2, \ldots) \mid X_i \in \{0, 1\}\},$$

[2]This is an obvious consequence of the Radon-Nikodym theorem.

[3]Note that constants are integrable with respect to this measure, so the bounded convergence theorem is just the dominated convergence theorem with a constant dominating function!

but it is unclear what $\mathcal{F}$ and $\mathbf{P}$ should be.

Define the events $A_{a_1,\ldots,a_n} = \{(X_1, X_2, \ldots) \in \Omega \mid X_1 = a_1, \ldots, X_n = a_n\}$ where $a_i \in \{0, 1\}$. Clearly, $\mathbf{P}(A_{a_1,\ldots,a_n})$ should be $1/2^n$. Consider the set

$$\mathcal{F}_0 = \{A_{a_1,\ldots,a_n} \mid n \in \mathbf{N}, a_i \in \{0, 1\}, i = 1, \ldots, n\} \cup \{\emptyset, \Omega\}.$$

Note that $\mathcal{F}_0$ is not *quite* a σ-algebra, for it is not closed under unions, countable or finite (it's closed under finite intersections and set subtraction, so it's actually a ring of sets). Recall that the Borel σ-algebra $\mathcal{B}$ was defined as the σ-algebra generated by finite open intervals along with $\mathbf{R}$ and $\emptyset$. It is the smallest σ-algebra containing all open sets, and the Lebesge measure of a Borel set B can be defined in terms of the measures of the Borel sets:

$$\mu(B) := \inf\{\sum_{j=1}^{\infty} \mu(B_j) \mid B \subset \bigcup_{j=1}^{\infty} B_j, \text{ where } B_j \in \mathcal{B} \text{ for every } j\}.$$

In a similar manner, we can use the sets $A_{a_1,\ldots,a_n}$ to generate a σ-algebra and a probability measure for the probability space Ω. This extension to a measure on a σ-algebra is described in Carathéodory's extension theorem.

Theorem 2.1 (Carathéodory)**.** *Let R be a ring of sets, and let μ be a measure on R, i.e. μ is non-negative and countably additive. Then μ extends to a measure on the σ-algebra generated by R.*

The Kolmogorov Extension Theorem rigorously proves a generalisation of the above approach.

Definition 2.2. Let T be an arbitrary set, and $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A *stochastic process* is a collection $\{X_t \mid t \in T\}$ of random variables X_t on the space $(\Omega, \mathcal{F}, \mathbf{P})$.

T is typically either countable and we have a sequence of random variables (e.g. coin flips), or T is an interval of the real line (an example would be $X_t = 1$ or 0 as to whether or not an atom decays at time t). However, the definition of a stochastic process allows T to be arbitrary. For brevity, we will denote the stochastic process $\{X_t \mid t \in T\}$ by X_T.

Let X_T be a stochastic process on $(\Omega, \mathcal{F}, \mathbf{P})$. This induces distributions in $\mathbf{R}^k$ for ever k: for every k-tuple $(t_1, \ldots, t_k) \in T^k$ of distinct elements, the random vector $(X_{t_1}, \ldots, X_{t_k})$ has distribution

$$\mu_{t_1,\ldots,t_k}(B) = \mathbf{P}((X_{t_1}, \ldots, X_{t_k}) \in B) \tag{1}$$

for every Borel set $B \in \mathcal{B}^k$. These distributions are the *finite-dimensional distributions* of the stochastic process.

Equation (1) implies two consistency facts about the finite dimensional distributions of X_T. For any $B = B_1 \times \cdots \times B_k \in \mathcal{B}^k$, consider a permutation π of $(1, 2, \ldots, k)$. The events

$$(X_{t_1}, \ldots, X_{t_k}) \in B_1 \times \cdots \times B_k \quad \text{and} \quad (X_{t_{\pi 1}}, \ldots, X_{t_{\pi k}}) \in B_{\pi 1} \times \cdots \times B_{\pi k}$$

are the same, so (1) implies that

$$\mu_{t_1,\dots,t_k}(B_1 \times \cdots \times B_k) = \mu_{t_{\pi 1},\dots,t_{\pi k}}(B_{\pi 1} \times \cdots \times B_{\pi k}) \tag{2}$$

for every finite subset $(t_1, \dots, t_k) \in T^k$. In other words, if $\mu_{t_1,\cdots,t_k}$ is the measure $\mu_1 \times \cdots \times \mu_k$ then $\mu_{t_{\pi 1},\dots,t_{\pi k}}$ is the measure $\mu_{\pi 1} \times \cdots \times \mu_{\pi k}$.

Next, notice that the events

$$(X_1, \dots, X_k) \in B_1 \times \cdots \times B_k \quad \text{and} \quad (X_1, \dots, X_k, X_{k+1}) \in B_1 \times \cdots \times B_k \times \mathbf{R}$$

are the same, so (1) would also imply that

$$\mu_{t_1,\dots,t_k}(B_1 \times \cdots \times B_k) = \mu_{t_1,\dots,t_k,t_{k+1}}(B_1 \times \cdots \times B_k \times \mathbf{R}) \tag{3}$$

for every t_{k+1} in $T - \{t_1, \dots, t_k\}$.

Kolmogorov's extension theorem lays the first step in the theory of infinite dimensional distributions by proving that the conditions imposed by (2) and (3) are sufficient to guarantee the existence of a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ on which there is a process X_T with the given finite dimensional distributions.

For a permutation π of $\{1, \dots, k\}$ and for any $m \geq k$ define

$$\psi_\pi(x_1, \dots, x_m) = (x_{\pi 1}, \dots, x_{\pi k}).$$

I claim that (2) and (3) together are equivalent to the fact

$$\mu_{t_1,\dots,t_k} = \mu_{u_1,\dots,u_m} \circ \psi_\pi^{-1} \tag{4}$$

whenever $\{t_1 \dots, t_k\}$ is a subset of $\{u_1, \dots, u_m\}$ and π is a permutation satisfying

$$(u_{\pi^{-1}1} \dots, u_{\pi^{-1}m}) = (t_1, \dots, t_k, u_{\pi^{-1}(k+1)} \dots, u_{\pi^{-1}m}).$$

For example, if $(u_1, u_2) = (u_1, t_1)$ and π swaps 1 and 2 then

$$\mu_{t_1}(B) = \mu_{u_1,u_2}(\mathbf{R}, B) = \mu_{u_1,u_2} \circ \psi_\pi^{-1}(B).$$

It's easy to see that (2) and (3) are special cases of (4) (choose $m = k$ and π appropriately to obtain (2); choose $m = k + 1$ and $\pi = 1$ to obtain (3)). Since ψ is a coordinate permutation followed by some number of projections, (2) and (3) imply (4) as well, so they are equivalent.

We are now ready to state and outline a proof of Kolmogorov's extension theorem.

Theorem 2.3. *Suppose that $\{\mu_{t_1,\dots,t_k} | t_j \in T, k \in \mathbf{N}\}$ is any collection of distributions such that for every k and for every subset $\{t_1, \dots, t_k, t_{k+1}\}$ of T,*

- $\mu_{t_1,\dots,t_k}(A_1, \dots, A_k) = \mu_{t_{\pi 1},\dots,t_{\pi k}}(A_{\pi 1}, \dots, A_{\pi k})$ *and*
- $\mu_{t_1,\dots,t_k}(A_1, \dots, A_k) = \mu_{t_1,\dots,t_k,t_{k+1}}(A_1, \dots, A_k, \mathbf{R})$

where π is an arbitrary permutation of $\{1, \dots, k\}$. Then there exists some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a stochastic process X_T on Ω such that $\mu_{t_1,\dots,t_n}$ are the finite dimensional distributions of X_T.

Proof. To keep the paper at a reasonable length, we will only prove select parts in detail. The proof is based on the proof given in [1], section 36 – the omitted details may be found there.

Step 1: First, we construct the space Ω and the σ-algebra $\mathcal{F}$. The probability measure is first defined on events in $\mathcal{F}$ that have some finite structure. These events form an algebra. The measure is then extended via Carathéodory's extension theorem to $\mathcal{F}$, using the same process by which the Lebesgue measure is extended from finite intervals to measurable sets

Let $\Omega = \mathbf{R}^T$, the set of functions $f : T \to \mathbf{R}$. Note that if $|T| = n < \infty$, $\mathbf{R}^T$ is just $\mathbf{R}^n$. Similarly, if $T = \{t_1, t_2, \ldots\}$, we have $\mathbf{R}^T = \{(x_1, x_2, \ldots) \mid x_j \in \mathbf{R}\}$.

For each $t \in T$, define $Z_t : \Omega \to \mathbf{R}$ by $Z_t(f) = f(t)$. The σ-algebra generated by the sets

$$\{f \in \mathbf{R}^T \mid Z_t(f) \in B\} = Z_t^{-1}(B),$$

where t ranges over T and B ranges over $\mathcal{B}$, will suffice. We will later see that the stochastic process Z_T will satisfy the requirements of the theorem.

Step 2: As alluded to above, we will first study the subclass $\mathcal{F}_0$ of sets of the form

$$\{f \in \mathbf{R}^T \mid (Z_{t_1}(f), \ldots, Z_{t_k}(f)) \in B\}$$

where $B \in \mathcal{B}^k$. These sets are called the *cylinders* of $\mathcal{F}$. Our first task is to show that $\mathcal{F}_0$ is an algebra of sets. Let $A = \{f \in \mathbf{R}^T \mid (Z_{t_1}(f), \ldots, Z_{t_k}(f)) \in B\}$ and $A' = \{f \in \mathbf{R}^T \mid (Z_{s_1}(f), \ldots, Z_{s_j}(f)) \in B'\}$.

- Because $\mathbf{R}^k - \mathcal{B}$ is a Borel set,

$$A^c = \{f \in \mathbf{R}^T \mid (Z_{t_1}(f), \ldots, Z_{t_k}(f)) \in \mathbf{R}^k - B\} \in \mathcal{F}_0.$$

- Let $(u_1, \ldots, u_m)$ be an m-tuple containing each t_i and s_i. Let $\widetilde{B} = \psi_{id}^{-1}(B)$ and $\widetilde{B}' = \psi^{-1}(B')$; $\widetilde{B}, \widetilde{B}' \in \mathcal{B}^m$. Equation (4) implies

$$A = \{f \in \mathbf{R}^T \mid (Z_{u_1}(f), \ldots, Z_{u_m}(f)) \in \widetilde{B}\}.$$

  We may write A' in a similar manner:

$$A' = \{f \in \mathbf{R}^T \mid (Z_{u_1}(f), \ldots, Z_{u_m}(f)) \in \widetilde{B}'\},$$

  so $A \cup A' = \{f \in \mathbf{R}^T \mid (Z_{u_1}(f), \ldots, Z_{u_m}(f)) \in \widetilde{B} \cup \widetilde{B}'\} \in \mathcal{F}_0$.

This shows that $\mathcal{F}_0$ is an algebra.

Step 3: We define a probability measure on the class of cylinders of $\mathcal{F}$,i.e. $\mathcal{F}_0$. If A is the cylinder $\{f \in \mathbf{R}^T \mid (Z_{t_1}(f), \ldots, Z_{t_k}(f)) \in B\}$, define

$$\mathbf{P}(A) = \mu_{t_1,\ldots,t_k}(B).$$

Since the Borel set B used in defining A is not necessarily unique, we must check the consistency of this definition. Not surprisingly, the consistency properties (2) and (3) are used for this. See [1], p. 489 for details.

By construction, the finite dimensional distributions of Z_T are precisely $\mu_{t_1,\ldots,t_k}$. Because $\mathbf{P}(\mathbf{R}^T) = \mu_t(\mathbf{R}) = 1$ for any $t \in T$, it remains to show that $\mathbf{P}$ is a countably additive measure on $\mathcal{F}_0$.

Step 4: We will first show that finite additivity holds. Instead of showing that $\mathbf{P}$ is countably additive on $\mathcal{F}_0$, we will use finite additivity to show that Kolmogorov's sixth axiom holds: if $A_n \in \mathcal{F}_0$ and $\bigcap_n A_n = \emptyset$ then $\mathbf{P}(A_n) \to 0$ as $n \to \infty$.

Consider disjoint cylinders A and A' given as in step 2. If $\widetilde{B}$ and $\widetilde{B}'$ were not disjoint, A and A' would also not be disjoint. Hence, $\widetilde{B} \cap \widetilde{B}' = \emptyset$, so

$$\begin{aligned}\mathbf{P}(A \cup B) &= \mu_{u_1,\ldots,u_m}(\widetilde{B} \cup \widetilde{B}') \\ &= \mu_{u_1,\ldots,u_m}(\widetilde{B}) + \mu_{u_1,\ldots,u_m}(\widetilde{B}') \qquad \text{(since } \mu_{u_1,\ldots,u_m} \text{ is assumed to be additive)} \\ &= \mathbf{P}(A) + \mathbf{P}(B)\end{aligned}$$

so $\mathbf{P}$ is finitely additive on $\mathcal{F}_0$.

Let $A_n \in \mathcal{F}_0$ be nested such that $\mathbf{P}(A_n) \geq \epsilon > 0$ for some fixed ϵ. We have to show that $\bigcap_n A_n$ is nonempty. Without loss of generality[4], assume that we can write

$$A_n = \{f \in \Omega \mid (Z_{t_1}(f), \ldots, Z_{t_n}(f)) \in B_n\}$$

for some sequence of parameters t_n and Borel sets $B_n \in \mathcal{B}^n$. Then $\mathbf{P}(A_n) = \mu_{t_1,\ldots,t_n}(B_n)$. For each n there exists a compact sets $K_n \subset B_n$[5] such that

$$\mu_{t_1,\ldots,t_n}(B_n - K_n) < \epsilon/2^{n+1}.$$

Define $\widetilde{A}_n = \{f \in \Omega \mid (Z_{t_1}(f), \ldots, Z_{t_n}(f)) \in K_n\}$, and $C_n = \bigcap_{k=1}^n \widetilde{A}_k$. Then $C_n \subset \widetilde{A}_n \subset A_n$, and

$$\begin{aligned}\mathbf{P}(A_n - C_n) &= \mathbf{P}(\bigcup_{k=1}^n (A_k - \widetilde{A}_k)) \\ &\leq \sum_{k=1}^n \mathbf{P}(A_k - \widetilde{A}_k) \\ &\leq \sum_{k=1}^\infty \mathbf{P}(A_k - \widetilde{A}_k) \\ &< \epsilon/2.\end{aligned}$$

This implies that $\mathbf{P}(C_n) > \epsilon/2$, so C_n is a sequence of nested, non-empty sets. By choosing an element $f^{(n)}$ in each C_n, we can use a diagonal argument to find a subsequence $f^{(n_i)}$ such that

$$x_{t_k} := \lim_{i \to \infty} Z_{t_k}(f^{(n_i)})$$

[4]This assumption is without loss of generality. See [1], p. 490.

[5]Recall from the fall semester that any set of finite measure could be approximated by a compact set. This is the corresponding fact for probability measures.

exists for every k. There is a function $f \in \Omega$ such that $Z_{t_k}(f) = x_{t_k}$ for every k. The compactness of K_k implies that

$$(Z_{t_1}(f), \ldots, Z_{t_k}(f)) = \lim_{i \to \infty} (Z_{t_1}(f^{(n_i)}), \ldots, (Z_{t_k}(f^{(n_i)}))$$

is in K_k, so $f \in \widetilde{A}_k \subset A_k$ for every k. This shows that $\bigcap_k A_k$ is nonempty as desired.

All of the prerequisite conditions of Carathéodory's theorem hold, so $\mathbf{P}$ extends to a measure on $\mathcal{F}$. □

3 Limitations of Kolmogorov's Theorem

The probability space and σ-algebra constructed in the previous section is not necessarily unique. In fact, for some applications, the σ-algebra $\mathbf{R}^T$ is inadequate. For instance, if T is an interval, the class of continuous functions from T to $\mathbf{R}$ is not measurable in this probability space.

This issue is dealt with on a somewhat case-by-case basis. Poisson processes and Brownian motion each have their own special construction of probability spaces with desirable properties. The concept of *separability* narrows the field of vision just enough to ensure that finite dimensional distributions with nice smooth properties extend to a stochastic process that also has nice smooth properties. If you are interested in this, see [1], section 38.

References

[1] Billingsly, Patrick. *Probability and Measure*, 3rd ed. John Wiley & Sons. 1995.

[2] Kolmogorov, Andrey N. *Foundations of the Theory of Probability*, 2nd English ed. Chelsea Publishing Company, New York. 1956.

[3] Shafer, G. and Vovk, V. *The origins and legacy of Kolmogorov's* Grundbegriffe. `http://www.probabilityandfinance.com/articles/index.html`.