Bertrand Paradoxes and Kolmogorov's Foundations of the Theory of Probability.

Christoforos Anagnostopoulos

October 19, 2006

Contents

Chapter 1

Introduction

1.1 The Paradoxes of Geometric Probability

Geometric probability is the informal study of the probabilities involved in geometric problems in which the random selection is one of points, lines or other geometrical objects in Euclidean space.

The field arose by extrapolation from the elementary, discrete case which had been the topic of earlier studies[1] to the continuous, geometric case. Hence, the most basic intuition of geometric probability was shaped, that

$$Pr(A) =_{df} \frac{\text{area covered by outcomes favorable to } A}{\text{area covered by all possible outcomes}} \tag{A}$$

where naturally one can replace 'area' by 'length' or 'volume' accordingly. This came only as a natural generalisation of the underlying principle of earlier mathematical studies of games of chance, namely that[2]

$$Pr(A) =_{df} \frac{\text{number of outcomes favorable to } A}{\text{number of possible outcomes}} \tag{B}$$

Setting the probability of an event equal to the proportion of outcomes favorable to the event out of all possible outcomes, be that proportion measured by 'length', 'area', 'volume' or just counting in the discrete case, is called the 'uniform' probability assignment.

[1] The study of games of chance goes back to 1494, when *Fra Luca de Paciolli* adresses the problem of the fair division of the stakes if a game is forced to stop midway (see [3]). However, as Professor Glenn Shafer informs us in [10], its systematic study essentially began with *Pierre de Fermat*'s and *Blaise Pascal*'s correspondence on that same matter. Continuous probability is hinted at by Isaac Newton in his study into the motions of the planet, around 1665. It was placed in the context of geometric problems some 70 years later, in 1773 when *Georges le Clerk Buffon* asked his famous *needle problem* in a lecture to the Paris Royal Academy of Sciences. He published the solution much later in 1777, in his *Histoire Naturelle, generalle et particuliére.*

[2] Formula (B) notably appears as the definition of probability in *Jacob Bernoulli*'s work, posthumously published in 1713 in the famous *Ars Conjectandi.*

Similarly, certain intuitions about the notion of *conditional probability*, usually amounting to an argument about the order of certain events or about a state of knowledge, were transferred over from the discrete to the continuous.

Both extrapolations turned out to be problematic.

Problematic Extrapolations - the Principle of Indifference

Formula (A) was immediately seen to be representation-dependent; a chord of a circle could be determined either by its two endpoints or by its midpoint, each parameterisation yielding a different value for the probability that the chord satisfies a certain property.

The same difficulty had already been recognised of (B), but was seen as surmountable in the discrete case since there always seemed to exist a fine-grained enough description with respect to which everyone agreed that elementary outcomes were symmetric, hence equiprobable. This criterion of 'symmetric elementary outcomes' came to be known as the 'Principle of Indifference'[3].

The mathematical content of this principle is that whenever a relabelling of the elementary outcomes leaves their probabilities unaffected (symmetry), the uniform distribution must be employed. This fails in the continuous case since clearly lengths, areas and volumes are preserved by some bijections but certainly not *all*. Symmetry in the elementary outcomes hence fails to uniquely identify a probability assignment. A growing list of geometric problems refusing to admit a general consensus representation loomed ahead.

Remark. The Principle of Indifference was so unequivocally employed in the discrete case that it was thought of as a *definition* of what probabilities are, not merely as a rule of thumb guiding an arbitrary choice of probability distribution. This explains why the failing of the Principle of Indifference, although brought about by simple mathematics, came as such a great shock to early thinkers.

Problematic Extrapolations: Conditional Probability

As for the calculus of conditional probabilities, theorists felt unsafe about which, if any, properties were preserved, since the definition itself could *not* transfer, designed as it was to model conditioning on events of nonzero probability only:

$$Pr(A \text{ given } B) =_{df} \frac{\text{number of outcomes favorable to both } A \text{ and } B}{\text{number of outcomes favorable to } B}$$

In contrast, in geometric probability it was often the case that we wished to condition upon an event of probability 0 (a great circle seen as a subset of the surface of the sphere, for instance), an operation that would force a division by zero in the definition above.

[3]This principle was first explicitly identified such by *Pierre-Simon Laplace* in his seminal work *Théorie Analytiques des Probabilités*, published in 1812. Laplace used the term *Principle of Insufficient Reason*, which was eventually abandoned for the term *Principle of Indifference*, coined much later by *John Maynard Keynes*. For further information see [10].

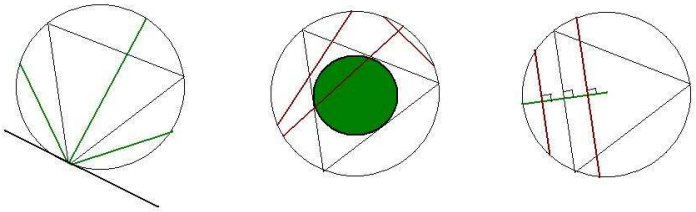

Figure 1.1: the Chord Paradox: from left to right, the three possible solutions to the question.

The Paradoxes

These difficulties were exemplified via concrete examples of problems that seemingly admitted no unique solution, of which we presently describe four representative specimens.

The Perfect Cube Paradox[4].

A precision tool factory produces iron cubes with edge length $\leq$ 1cm. What is the probability that the face area of a randomly selected cube is $\leq 0.25\text{cm}^2$? Observing that each edge length $\leq$ 1cm corresponds uniquely to a face area $\leq$ 1cm, we can recast this problem more abstractly as follows.

Let $y = x^2$. For any random choice of $x \in [0, 1]$, essentially a choice of $y \in [0, 1]$ has also occurred. However, the probability that x lies in $[0, 1/2]$ is $1/2$, whereas the probability that y lies in $[0, 1/4]$ is $1/4$ - but the two probabilities ought to be equal since the two events are equivalent.

The Chord Paradox[5].

Consider a disk on the plane with an inscribed equilateral triangle. What is the probability that a chord chosen at random be longer than the side of an inscribed equilateral triangle?

The position of the inscribed triangle relative to the chord chosen is clearly irrelevant to the problem. So we may as well assume that one of the two points of the chord is fixed on one of the vertices of the inscribed triangle. Now, the

[4]The underlying idea of this argument has been severally phrased by various authors at various times. The transformation $y = x^2$ on $[0, 1]$ was used by *Emile Borel* and *Henri Poincaré* (see [14]), whereas $y = 1/x$ on $\mathbb{R}^+$ was instead used by *Johannes Von Kries*. The anecdotal recasting in terms of a perfect cube factory is due to *Bas C. Van Fraassen*, in [13].

[5]This paradox is most commonly known as *Bertrand's paradox*. Van Fraassen in [13] informs us it was first described by *Joseph Bertrand* in his 1888 textbook *Calcul des Probabilités*, along with the Paradox of the Great Circle, that follows. Bertrand was a well-known probabilist and his views on the ill-defined nature of certain continuous probability problems were very influential, which explains why paradoxes of the type encountered in this section are often called *Bertrand-type paradoxes*.

two edges of the triangle emanating from that point trichotomize the tangent to that point, so 1/3 of the outcomes will result in a chord longer than the side of the triangle.

However, a chord is also completely determined by its midpoint. Chords whose length exceeds the side of an inscribed equilateral triangle have their midpoints inside a smaller circle with radius equal to 1/2 that of the given one. The set of favorable midpoints covers an area 1/4 of the original disk, which also defines the proportion of favorable outcomes as 1/4, not 1/3.

Alternatively, by rotational symmetry, we may fix the radius that the midpoint of the randomly selected chord will lie on. Then the proportion of favorable outcomes is precisely all points on the radius that are closer to the center than half the radius, so it is neither 1/3, nor 1/4, but 1/2.

The Paradox of the Great Circle[6].

Pick two points on the sphere at random. What is the probability that they lie within $10'$ of each other? By symmetry we assume that the first point is fixed on the North Pole of the sphere. We then calculate the proportion of the sphere's surface that lies within $10'$ of the North Pole. This is 2.1×10^{-6}.

We may however observe that there exists a unique great circle that connects the second randomly selected point with the North Pole. Moreover, by rotational symmetry, no great circle has more chances of being selected than any other. Therefore, we may assume we know the great circle that connects the two points. We have now reduced the original problem to one of picking one point on a given great circle. The answer to the original question can hence be found by calculating the proportion of the length of the great circle that lies within $10'$ of the North Pole, which is of course $10/(180 \cdot 6) \approx 9.3 \times 10^{-3}$, not 2.1×10^{-6}.

Buffon's Needle Problem[7].

Assume that a large number of parallel lines 10cm apart are drawn on the floor and a needle is dropped. What is the probability that a needle 5cm in length intersects with one of the lines? Clearly the needle will intersect at most one line. The two quantities of interest can be therefore seen to be the distance d of the needle's tip to the line that it is nearest to and the angle θ that the needle forms with that line. It is natural to assume that the two are independent and chosen at random. Moreover, the favorable outcomes are precisely those for which $d \leq \sin\theta$. Now d varies from 0 to 5 and θ from 0 to 2π. The proportion of this area that satisfies the equation $d \leq \sin\theta$ is equal to $1/\pi$, which also gives the required probability. Until this day, Buffon's solution is recognised as

[6]This was published in *Bertrand*'s *Calcul des Probabilités* (see previous footnote) and was revisited by *Borel* shortly afterwards. It is the only problem of geometric probability that Kolmogorov himself adresses in the Grundbegriffe, where he refers to it as *a Borel paradox*([6, p. 50-51]). Currently it is mostly known as the *Borel-Kolmogorov* paradox.

[7]See Footnote 1.

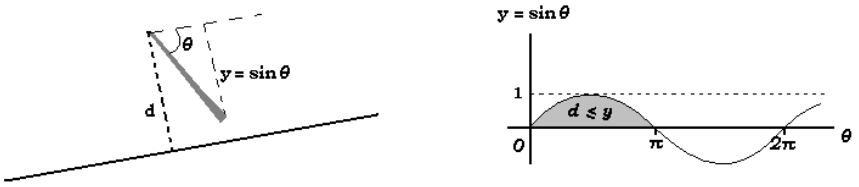

Figure 1.2: Buffon's Needle Problem: on the left, the two quantities of interest, d and $y = \sin\theta$. On the right, the calculation of the area representing favorable outcomes, under the (θ, d) parameterisation.

correct and fully agrees with experimental data of needle throwing[8]. However, we need only reparameterise the problem to represent $\sin\theta$ by y to get a different answer.

1.2 The Call for Foundations

The early shock caused by the heterogeneity of geometric and discrete probability had by 1933 been replaced by a puzzled enthusiasm, since probabilities had by then been employed succesfully in a wide range of problems: discrete games of chance, continuous modelling of prices in the stock exchange, stability of planetary orbits, properties of decimal expansions of real numbers and notably statistical and quantum mechanics. Each application would often come with its own interpretation of what 'probabilities' meant, contributing towards the confusion surrounding that question which prevailed amidst philosophers and mathematicians. On the other hand, each application would also contribute to the growing toolbox of theorems and methods for probability theory, bridging gaps, creating more abstract versions of existing results and overall adding to the spreading belief that Kolmogorov succinctly expressed in a 1929 publication:

> [...] one gains the impression that the formulas of the calculus of probability express one basic group of mathematical laws of the most general kind[9].

Indeed the laws of probability by 1933 were conceived as axioms, to be satisfied by various interpretations, not as truths that must be strained to hold universally to a variety of settings[10]. Kolmogorov in his introduction to the

[8] Buffon's solution can be trusted to the extent that it can be used as a way to obtain arbitrarily good approximations to π.

[9] The quote is borrowed from [14, p.21].

[10] (This discussion is a synthesis of [14], [11] and [10]). Already in 1901, in Georg Bohlmann and Ladislaus Von Bortkiewicz's entry on *probability* in the *Encyklopadie der mathematischen Wissenschaften* the addition and multiplication theorems were stated as "axioms". David Hilbert called for an axiomatisation of probability one year before that, in 1900. *Rudolf Laemmel*'s dissertation in 1904 attempted to offer precisely such an axiomatic basis for re-

Grundbegriffe puts it as follows ([6, p.1]):

> The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by which these relations are to be governed, all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their relations. [...] Every axiomatic (abstract) theory admits, as is well known, of an unlimited number of concrete interpretations besides those from which it was derived.

This reconceptualisation gave much-needed breathing space to mathematicians involved in the theory of probability. In its light, the development of the mathematics of probability could proceed uninhibited by the persistent conundrum of what probabilities 'really meant' and the paradoxes could be seen as ill-advised *applications* of probability theory - much less to the despair of the mathematician than *inconsistencies* within the theory.

No More Paradoxes?

In particular with respect to geometric probability and Bertrand-type paradoxes, Kolmogorov's foundations broke the matter down as follows:

Step A. A formal construct is offered, now known as a *probability space*, that is meant to model any 'random experiment', ie any statement of the form 'pick $x \in X$ at random'. This reductive imperative underlies the entire measure-theoretic conception of probability theory:

> Any well-posed question about the probability of an event $\mathcal{A}$ must be in unique correspondence to a question about the probability of a set A that is represented in a specified *probability space*.

An important step in the right direction is that probabilities are now thought of as assignments on *sets of elementary outcomes*, as opposed to assignments on the elementary outcomes themselves. In the discrete case, the two approaches are equivalent, since assignments on the singletons extend uniquely to assignments on arbitrary sets by the properties of probability. In the continuous case, however, this is not so - this being yet another way to observe the failing of the Principle of Indifference here.

---

ducing probability theory to set theory, whereas a similar attempt was made by *Ugo Broggi* shortly afterwards in 1907, in which measure theory was also employed. The choice of axioms and definitions in these two works were however not in consonance with the viewpoint that was eventually adopted, as expressed in the Grundbegriffe. In contrast, Evgeny Slutsky's publication in 1922 of *On the question of the logical foundation of the theory of probability* was praised by Kolmogorov as "the first to give the right picture of the purely mathematical content of probability theory". In it Slutsky suppressed mention of probabilities and equally likely cases altogether in favor of a theory of *valences*, using additivity with respect to partitions, while he referred to the various schools of probability theory as *interpretations* of the theory of valences.

Step B. One then proceeds to establish that the operative notions of informal probability talk can be formalised and indeed extended in the context of probability spaces. Such notions involve functions from one set of elementary outcomes to another, expectations (or mean values), independence, joint consideration of several random experiments and, notably, conditional probability with respect to an event of probability 0.

Step C. Finally, whenever two formal calculations yield different answers, this can only be because the underlying probability spaces are different (since, by definition, the probability space specifies uniquely the probabilities of all events one is allowed to reason with). The question of which probability space, if any, is the one that 'truly' models this particular question in geometric probability is one of *applicability* and is investigated separately from the mathematics.

The Mathematics of the Grundbegriffe

This dissertation will provide a modern viewpoint of precisely the mathematics laid out in *Step B* above. In a sense, it was precisely the fact that the mathematics worked so nicely that dictated the choice of a probability space in *Step A* as a fundamental notion. It will also become apparent as the dissertation proceeds that the mathematics we will use rest on nontrivial developments, recent at the time of the Grundebgriffe, in *measure theory*[11]. We take great care throughout the main text to make explicit any such dependencies, even when Kolmogorov fails to do so. We do this to emphasize the point that it would have been exceptionally hard for the mathematics community to provide foundations for even relatively simple fragments of probability theory any earlier than the 1920's, in the absence of the general framework of measure theory.

Even with the groundwork laid and the mathematical community already convinced of the fundamental relationship between probability theory and measure, the Grundbegriffe remains an important work not only as a work of consolidation, but also because of certain novel technical contributions that in a sense provided the *coup de grâce* to any doubts as to the value of Kolmogorov's axiomatisation. These were

- the introduction of infinite-dimensional probability spaces,

[11] (This discussion is a synthesis of [14], [11] and [10]). The theory of measure arguably starts with Borel's publication of *Leçons sur la théorie des fonctions*, wherein he describes his theory of *Borelian content*, a generalisation of length on the real line. This is followed by *Henri Lebesgue*'s 1904 *Leçons sur l'intégration et la recherche des fonctions primitives*, in which the theory of *Lebesgue measure*, a broad generalisation of that of Borelian content is introduced. Shortly afterwards, a theory of integration suitable for Lebesgue measure started developping by Lebesgue himself but also by others, with the *Radon integral* appearing in *Johann Radon*'s *Theorie und Anwednungen der absolut additiven Mengenfunktionen.* In 1914, *Constantin Carathéodory* published his *Über das lineare Mass von Punktmengen*, in which his notion of outer measure and generalised integration appears, followed by the abstract *Fréchet* integral, featured in *Maurice Fréchet*'s *Sur l'intégrale d'une fonctionelle étendue à un ensemble abstrait.* Finally, Kolmogorov's notion of conditional probability rested on the *Radon-Nikodym* theorem, stated in its fully abstract form in 1930 by *Otto Nikodym* in his *Sur une généralisation des intégrales de M.J.Radon.*

- the demonstration of techniques for the differentiation and integration of expectations with respect to a parameter,
- the precise definition of a notion of conditional probability with respect to events of probability 0.

In our discussion in the main text we overview the latter two contributions in detail, since they pertain to questions of geometric probability. The former contribution does not and is hence omitted.

Afterword

There are several standpoints as to whether it is possible to venture beyond Kolmogorov's *Step C* so as to also resolve via mathematics the question of *applicability* of probability theory to geometric problems, in the sense of providing formal or at least semi-formal criteria as to the appropriate choice of probability space for each problem in geometric probability. Some think it impossible, in effect condemning geometric probability problems to be intrinsically ambiguously phrased. This would make Kolmogorov's formalism a somewhat pyrrhic victory over the paradoxes, as is succinctly explained in the following quote by the philosopher of science Mark Van Fraassen ([13, p.305]):

> After all, if we were told as part of the problem which parameter should receive a uniform distribution, no such Principle [as that of Indifference] would be needed. It was exactly the function of the Principle [of Indifference] to turn an incompletely described physical problem into a definite problem in the probability calculus.

Indeed, others think more can be achieved and that the Principle of Indifference can be replaced by more sophisticated semi-formal criteria. However, to do so, one must investigate and take a position as to the nature of the object of study of geometric probability. Is it purely mathematical, is it an idealisation of physical objects or is it a form of logical calculus?

To discuss such questions thoroughly falls beyond the scope of this dissertation. However, in an afterword to Chapter 3, we will refer the interested to reader to one influential school of thought that has attempted, with some success, to further Kolmogorov's revisionist programme so that it can accommodate a much larger fragment of geometric probability talk.

Chapter 2

The Grundbegriffe

This Chapter forms the main part of this dissertation, where we review the mathematics underlying the measure-theoretic conception of probability theory. We present two intertwined narratives, one faithful to modern standard practice, the other faithful to Kolmogorov's original mode of exposition, as captured in the Grundbegriffe. Whereas the former is more streamlined, the latter is often more revealing of basic intuitions.

This Chapter is subdivided as follows. In Section 2.1, we introduce and familiarise ourselves with the basic notions: *probability spaces*, *independence*, *elementary conditional probability* and *probability functions*.

Section 2.2 plays the role of infrastructure works. We provide and discuss certain measure-theoretic results that make possible the construction of interesting probability spaces, in particular over countable domains and over $\mathbb{R}^n$. The operative results here are *Carathéodory's Extension Theorem* and *Fubini's Product Measure Theorem*.

Section 2.3 uses the results of Section 2.2 to provide concrete *examples of probability spaces* modelling familiar problems. We do not advance the theory here at all; we merely toy around with the notions already defined.

Section 2.4 ventures ahead to the study of *random variables* and their *distribution functions*. We establish the correspondence between distribution functions and probability measures over the Borel sets of $\mathbb{R}^n$ and argue for the singular role of Lebesgue measure in the foundations of probability theory.

Section 2.5 introduces the notion and properties of *mathematical expectations*, an application of *Lebesgue integration*. We take some time to investigate a somewhat neglected Theorem of Kolmogorov's and apply it to the puzzling experiment of *randomly spattering a wall with paint*.

Finally, in Section 2.6 we motivate, introduce and discuss *Kolmogorov's notion of conditional probability*.

In certain sections, we will adopt a more discursive style so as to offer some intuition, put a certain result in historical perspective or perhaps give an idea

of an omitted proof. We take care to include the qualification "a discussion" in the title of such sections.

2.1 Probability Spaces

The Axioms of Probability Spaces

Our basic object throughout this dissertation is a triple $(\Omega, \mathfrak{F}, P)$, where

- $\Omega = \{\xi_i \mid i \in I\}$ is the set of *elementary outcomes* or *outcomes*,
- $\mathfrak{F} \subseteq \mathcal{P}(\Omega)$ is the set of *random events* or *events*,
- $P : \mathfrak{F} \to \mathbb{R}$ is a real-valued set function that assigns probabilities to events.

One has to get used to this vocabulary whereby *events are sets of outcomes*. Our aim is not merely to assign probabilities to individual outcomes, but rather to be able to talk of the probability that any one out of of a given *set* of outcomes occurs (this set could be an interval of values, for instance). On the other hand, single outcomes are still represented by singletons.

Definition 1. We call the tuple $(\Omega, \mathfrak{F}, P)$ (or the tuple $(\mathfrak{F}, P)$ whenever Ω is clear from the context) a **probability space** if it satisfies the following axioms:

I. $\mathfrak{F} \subseteq \mathcal{P}(\Omega)$ is a family of subsets of Ω closed under complements, countable unions and countable intersections,

II. $\Omega \in \mathfrak{F}$,

III. P is a function from $\mathfrak{F}$ to $\mathbb{R}^+$, the set of positive real numbers,

IV. $P(\Omega) = 1$,

V. if $A, B, A \cup B \in \mathfrak{F}$ and $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$,

VI. if $A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$ is an infinite descending chain of events of $\mathfrak{F}$ with an empty intersection $\bigcap_{n\in\mathbb{N}} A_n = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$.

In modern terms, Axioms I and II can be jointly stated as follows:

Definition 2. A family $\mathfrak{F}$ of subsets of some set Ω is said to be a **σ-algebra** over Ω iff it contains Ω and is closed under complements and countable unions.

In some cases, however, a weaker set of conditions on $\mathfrak{F}$ is all we need:

Definition 3. A collection of subsets of Ω is an *algebra* over Ω whenever it contains Ω and is closed under pairwise unions, differences and intersections. An algebra is therefore finitely closed, as opposed to being countably closed.

It makes sense in the first couple of sections to keep track of our assumptions, so that the role of each axiom be clearly understood by the reader. In particular, it will be important to make a note of which results can go through in the absence of Axiom VI. This motivates the following definition:

Definition 4. Let $\mathfrak{F}$ be an algebra over Ω. If P satisfies Axioms III-V over $\mathfrak{F}$ then it is called *finitely additive*.

The objects of interest therefore at this early stage are all tuples $(\Omega, \mathfrak{F}, P)$ where $\mathfrak{F}$ is either an algebra or a σ-algebra and P is certainly finitely additive, but may or may not satisfy Axiom VI. Eventually[1], however, we will only be interested in probability spaces, ie tuples that satisfy the whole of Definition 1.

Some immediate corollaries

We now derive several elementary properties of probabilities that flow from finite additivity. Easy set-theoretical results are stated without proofs.

Proposition 1. *Consider $(\Omega, \mathfrak{F}, P)$ where $\mathfrak{F}$ is an algebra and P finitely additive, ie satisfies Axioms III-V. Then, if $A, B, A_1, A_2, ... \in \mathfrak{F}$, we have*

(a) $P(A) + P(A^c) = 1$.

(b) $P(\emptyset) = 0$.

(c) *If $A_1, ..., A_n$ are pairwise disjoint[2], then $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.*

(d) *If $A_1, ..., A_n$ are pairwise disjoint and $\bigcup_{i=1}^n A_i = \Omega$, then $\sum_{i=1}^n P(A_i) = 1$.*

(e) *If $A \subseteq B$, then[3] $P(B \setminus A) = P(B) - P(A)$.*

(f) *If $A \subseteq B$, then $P(A) \leq P(B)$.*

(g) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. For Claim (a), we observe that $P(A \cup A^c) = 1$ by Axiom IV since $A \cup A^c = \Omega$. But A^c is in $\mathfrak{F}$ by finite closure of $\mathfrak{F}$. Since A and A^c are mutually disjoint, $P(A \cup A^c) = P(A) + P(A^c) = 1$ by Axiom V. Claim (b) then follows from (a) by setting $A =_{df} \Omega$. Claim (c) follows from a trivial inductive application of Axiom V, whereas claim (d) follows from (c) using $P(\Omega) = 1$. For Claims (e) and (f), we observe that A and $B \setminus A$ are disjoint, so:

$$\begin{aligned} P(B) = P(A \cup (B \setminus A)) &= P(A) + P(B \setminus A), \text{ by Axiom V, yielding (e)} \\ &\geq P(A), \text{ since } P \geq 0, \text{ by finite closure of } \mathfrak{F}, \text{ yielding (f)} \end{aligned}$$

[1] Kolmogorov himself initially requires only that $\mathfrak{F}$ be an algebra and P finitely additive, then introduces Axiom VI and countable closure of $\mathfrak{F}$ at a later stage. We thoroughly explain his approach and the reasons behind it later, in Section 2.2.

[2] Kolmogorov adopts a notational convention that distinguishes between the union of $A_1, ..., A_n$ when we know these to be disjoint ($\sum_i A_i$) than when we do not ($\bigcup_i A_i$). In this way, he turns the additivity of P into the statement that P commutes with the Σ operator. This method of making assumptions implicit in the notation is recurrent in Kolmogorov but not standard today.

[3] Kolmogorov uses $B - A$ to denote the difference of the two sets. See previous footnote.

Finally for Claim (g), observe that:

$$\begin{aligned} P(A \cup B) &= P(A \cup (B \setminus A)) = P(A) + P(B \setminus A), \text{ by Axiom V} \\ &= P(A) + P(B \setminus (A \cap B)), \text{ since } B \setminus A = B \setminus (A \cap B) \\ &= P(A) + P(B) - P(A \cap B), \text{ by (e)} \end{aligned}$$ □

Countable Additivity

We have made no use of Axiom VI in the proof of Proposition 1. This is because we have been concerned with finite collections of events only. It is an easy observation that if finite collections is all we can get out of $\mathfrak{F}$, Axiom VI is redundant.

Proposition 2. *If $\mathfrak{F}$ is a finite family of subsets of Ω and P is finitely additive over $\mathfrak{F}$, then P also satisfies Axiom VI over $\mathfrak{F}$.*

Proof. Any descending chain $A_1 \supseteq A_2 \supseteq ...$ has to involve finitely many strict inclusions (since each strict inclusion introduces a new element of the finite set $\mathfrak{F}$). Therefore it eventually becomes constant: $A_N = A_{N+1} = ... = \bigcap_n A_n = \emptyset$. This proves that $\lim_{n\to\infty} P(A_n) = P(A_N) = P(\emptyset)$, as required. □

Remark. The cardinality of Ω does not come into play in the above proof.

When we deal with infinite families of events, Axiom VI becomes necessary to ensure that P is well-behaved in that it commutes with certain natural limit operations. The limit operation we have so far considered is very specific (the infimum of a descending chain with empty intersection). It is however sufficient to yield the same property for P with respect to other natural limit operations:

Proposition 3. *Let $\mathfrak{F}$ be a σ-algebra over Ω and P be finitely additive over $\mathfrak{F}$. Then the following are equivalent, for sets A_i, B_i, C_i, D_i all in $\mathfrak{F}$:*

(a) [Axiom VI] *If $A_1 \supseteq A_2 \supseteq ...$ and $\bigcap_i A_i = \emptyset$, then $\lim_{i\to\infty} P(A_i) = 0$.*

(b) *If $B_1 \supseteq B_2 \supseteq ...$, then $\lim_{i\to\infty} P(B_i) = P(\bigcap_i B_i)$.*

(c) *If $C_1 \subseteq C_2 \subseteq ...$, then $\lim_{i\to\infty} P(C_i) = P(\bigcup_i C_i)$.*

(d) *If $D_1, D_2, ...$ are pairwise disjoint, then $\sum_i P(D_i) = P(\bigcup_i D_i)$.*

Proof. Our method of proof will be to establish that $a \Leftrightarrow b$, $b \Leftrightarrow c$ and $d \Leftrightarrow a$. Firstly, (a) is a special case of (b), so the direction (b) $\Rightarrow$ (a) is trivial. We now show that (a) $\Rightarrow$ (b). Denote by B the intersection of the chain in (b). Then $P(B_i) = P(B_i \setminus B) + P(B)$. But now consider the following chain:

$$(B_1 \setminus B) \supseteq (B_2 \setminus B) \supseteq ...$$

Its intersection is empty, so we can apply (a) to obtain:

$$\lim_n (P(B_n) - P(B)) = \lim_n P(B_n) - P(B) = 0$$

as required. We have shown $(a) \Leftrightarrow (b)$ and (b) $\Leftrightarrow$ (c) is trivially obtained by taking complements. We presently show that $(a) \Rightarrow (d)$. Let:

$$D =_{df} \bigcup_n D_n, \qquad D_1, D_2, ... \in \mathfrak{F}, \text{ with the } D_i\text{'s pairwise disjoint}$$

Now let $A_m = \bigcup_{n>m} D_n$. Clearly the A_i's form a descending chain. Moreover their intersection is empty, since:

$$[\forall m,\, x \in A_m] \;\Leftrightarrow\; [\forall m, \exists n > m,\, x \in D_n] \;\Leftrightarrow\; x \text{ is in infinitely many } D_n\text{'s}$$

which contradicts the disjointness of the D_n's. So by an application of (a):

$$\lim_{m\to\infty} P(A_m) = 0$$

This completes the proof, since $P(D) = P(D_1) + ... + P(D_m) + P(A_m)$ for each m, by Axiom V, which yields that $P(D) = \sum_m P(D_m)$ by taking the limit as m tends to infinity of each side.

Finally, we show that $(d) \Rightarrow (a)$. Let $A_1 \supseteq A_2 \supseteq ...$ be a descending chain with empty intersection. Clearly then the complements of the A_i's form an ascending chain whose union is the whole of Ω. We can of course turn their union into a union of disjoint sets, by letting:

$$D_1 = A_1^c \;\text{ and for } i > 1,\; D_i =_{df} (A_i^c \setminus A_{i-1}^c) = (A_{i-1} \setminus A_i)$$

We can then apply (d) and Proposition 1 to obtain that $\lim_{n\to\infty} P(A_n) = 0$:

$$\begin{aligned}
1 = P(\Omega) &= \sum_i P(D_i), \text{ by applying (d) and using } \bigcup_i D_i = \Omega \\
&= P(A_1^c) + \sum_{i>1} P(A_{i-1} \setminus A_i), \text{ by definition of } D_i \\
&= P(A_1^c) + \sum_{i>1} (P(A_{i-1}) - P(A_i)), \text{ by Prop 1, also using } A_{i+1} \subseteq A_i \\
&= P(A_1^c) + \lim_{n\to\infty} (P(A_1) - P(A_n)), \text{ since all other terms cancel out} \\
&= P(A_1^c) + P(A_1) - \lim_{n\to\infty} P(A_n) = 1 - \lim_{n\to\infty} P(A_n), \text{ as required.} \qquad \square
\end{aligned}$$

Properties (b) and (c) are sometimes known as *monotone convergence* properties. Property (d) is known as the property of *countable additivity*. A probability assignment is then finitely additive iff it satisfies Axioms III-V and countably additive iff it satisfies Axiom VI as well - in this latter case, it can also be referred to as a **probability measure**[4]:

Remark. The tuple $(\Omega, \mathfrak{F}, P)$ is a probability space if and only if $\mathfrak{F}$ is a σ-algebra containing Ω and P a probability measure over $\mathfrak{F}$.

[4]In measure theory, a set function $\mu : \mathfrak{F} \to \mathbb{R}^+$ defined on a family $\mathfrak{F}$ of subsets of Ω that satisfies countable additivity is called a *measure* iff $\mu(\emptyset) = 0$, a *finite* measure iff $\mu(\Omega) < \infty$ and a *probability measure* iff $\mu(\Omega) = 1$.

Elementary Conditional Probability

We now introduce some new notions. We start with that of *elementary conditional probability*. We presently give its formal definition, immediately followed by some motivation in the form of Proposition 4.

Definition 5. [Elementary Conditional Probability] The *elementary conditional probability* $P_A(B)$ *of* B *given* A, where $P(A) > 0$, is given by:

$$P_A(B) =_{df} \frac{P(A \cap B)}{P(A)} \tag{2.1}$$

or, equivalently, it can be given as the unique solution of:

$$P(A \cap B) = P(A)P_A(B) \tag{2.2}$$

Remark. Arguably (2.2) has a mild advantage over (2.1), since when $P(A) = 0$ it merely fails to produce a unique solution, rather than force an illegal operation[5].

Proposition 4. *Let* $A \in \mathfrak{F}$ *with* $P(A) > 0$ *and* $(\Omega, \mathfrak{F}, P)$ *be a probability space. Then* $(\Omega, \mathfrak{F}, P_A)$ *is also a probability space.*

Proof. With immediate proofs:

$$P_A(B) \geq 0, \quad P_A(\Omega) = 1, \quad P_A(B \cup C) = P_A(B) + P_A(C) \quad (B, C \text{ disjoint})$$

That P_A satisfies Axiom VI is nearly as direct. If $(A_n : n \in \mathbb{N})$ is a descending chain of events with empty intersection, so is $(A \cap A_n : n \in \mathbb{N})$, hence:

$$\begin{aligned} \lim_{n\to\infty} P_A(A_n) &=_{df} \frac{1}{P(A)} \lim_{n\to\infty} P(A \cap A_n), \text{ by Def 5 and extracting } 1/P(A) \\ &= 0, \text{ by Axiom VI as it applies to } \lim_{n\to\infty} P(A \cap A_n). \end{aligned}$$ □

Notation. The modern notation for $P_A(B)$ is $P(B \mid A)$. Observe that Kolmogorov's notation $P_A(B)$ is much more compact in that it readily suggests a nice subscripted piece of notation for the resulting distribution, P_A.

Observe that Proposition 4 together with (2.2) imply that P_A is in fact the unique probability measure P' on $\mathfrak{F}$ such that $P'(B)$ is always proportional to $P(A \cap B)$. This sheds some light into the meaning of elementary conditional probability; conditioning on an event A has the effect of reevaluating the probability of any other event B to take into account as a given fact that A has also occurred. Let us now derive some simple results about conditional probabilities.

[5]It also perhaps mirrors best the intuition of early workers in probability theory, who took $P(A) = 0$ to mean that A was impossible (a view which was later abandoned with the introduction of continuous probability). Since the probability of B conditional on A was then essentially understood as a hypothetical statement ("if A has happened, then the probability of B is x"), A being impossible meant that the hypothetical statement had an impossible antecedent, in which case its consequent (and hence also the value of the conditional probability of B) could be anything whatsoever.

Proposition 5. *Let $X, A_1, A_2, ..., A_n \in \mathfrak{F}$ and $P(A_i) > 0$ for all i. Then:*

(a) Multiplication Theorem:

$$P(A_1, A_2, ..., A_n) = P(A_1)P_{A_1}(A_2)P_{A_1A_2}(A_3)...P_{A_1...A_{n-1}}(A_n)$$

(b) Total Probability Theorem: *if the A_i's are disjoint and $\bigcup_i A_i = \Omega$, then*

$$P(X) = P(A_1)P_{A_1}(X) + ... + P(A_n)P_{A_n}(X)$$

(c) Bayes' Theorem: *if the A_i's are disjoint, $\bigcup_i A_i = \Omega$ and $P(X) > 0$, then*

$$\forall i,\ P_X(A_i) = \frac{P(A_i)P_{A_i}(X)}{P(A_1)P_{A_1}(X) + ... + P(A_n)P_{A_n}(X)}$$

Proof. Claim (a) follows by trivial induction on (2.2). Moreover, clearly:

$$X = \bigcup_{i=1}^{n} X \cap A_i, \qquad \text{where the } X \cap A_i\text{'s are disjoint}$$

so Claim (b) follows by additivity and (2.2) applied on each $X \cap A_i$. Finally,

$$P_X(A_i) =_{df} \frac{P(X \cap A_i)}{P(X)} = \frac{P(A_i)P_{A_i}(X)}{P(X)}$$

which yields precisely (c) if we substitute in the expression in (b) for $P(X)$. □

Remark. Result (c) is known as *Bayes' Theorem* and is the cornerstone of the theory of Bayesian Inference. Viewed in a context of hypothesis testing, Bayes' Theorem acts as a *learning* scheme (otherwise known as a *belief revision* scheme), whereby we interpret the pairwise disjoint sets A_i as mutually exclusive competing hypotheses about a certain experiment and X as an experimentally observed event (ie, a dataset). Then, for each i, the quantity $P_X(A_i)$ measures how probable hypothesis A_i has become given the occurrence of the data X. Bayes' theorem allows us to express this quantity in terms of:

- how likely it is under each hypothesis for data X to come up (the $P_{A_i}(X)$'s).
- how likely each hypothesis was prior to the observation X (the $P(A_i)$'s).

This is very intuitive; a quantitative version of the sort of argument common sense would produce. And yet Bayes' theorem constitutes a major point of dispute between Bayesian probability theorists and measure-theoretic probability theorists, since Bayesians often (ab)use the aforementioned interpretation by applying the theorem in situations where it is not clear what the underlying probability space is, or even whether there is one.

Independence

As we have seen in the introduction, a concept closely related to conditional probabilities is that of independence.

Definition 6. Two events A_1, A_2 are *independent* iff:

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

Proposition 6. *Two events A_1, A_2 of positive probability are independent iff:*

$$\begin{aligned} & P_{A_1}(A_2) = P(A_2) \\ \textit{or equivalently} \quad & P_{A_2}(A_1) = P(A_1) \end{aligned}$$

Proof. Follows directly from Definition 5. □

Remark. The semantics of independence and conditional probability are then intertwined: to condition on a certain event A is to assume it has happened - if to do so leaves the probability of a certain event B unaffected, then the two are said to be independent.

We now generalise our definition to apply to collections of events larger than merely a pair. There are of course two ways to start:

Definition 7. Events $A_1, ..., A_n$ are *(mutually) independent* iff the following holds for all $m \leq n$ and $1 \leq i_1 < ... < i_m \leq n$:

$$P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_m}) = P(A_{i_1})P(A_{i_2})...P(A_{i_m})$$

Events $A_1, A_2, ..., A_n$ are *pairwise independent* iff:

$$P(A_i \cap A_j) = P(A_i)P(A_j), \text{ for all } i \neq j$$

In the case of independence, unlike that of disjointness, 'mutually' is strictly stronger than 'pairwise':

Proposition 7. *Pairwise independence does not imply mutual independence.*

Proof. It is easy to offer a counterexample. This one is due to S.N. Bernstein[6]. Let:

$$\begin{aligned} & \Omega =_{df} \{\xi_1, \xi_2, \xi_3, \xi_4\}, \mathfrak{F} =_{df} \mathcal{P}(\Omega) \\ & P(\{\xi_i\}) = 1/4, \text{ from which we can derive all other values,} \\ & A =_{df} \{\xi_1, \xi_2\},\ B =_{df} \{\xi_1, \xi_3\},\ C =_{df} \{\xi_1, \xi_4\} \end{aligned}$$

We then have that $P(A) = P(B) = P(C) = 1/2$. Now observe that:

$$\begin{aligned} & P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4 = (1/2)^2 \text{ (pairwise independence)} \\ & \text{but } P(A \cap B \cap C) = P(\{\xi_1\}) = 1/4 \neq (1/2)^3 \text{ (mutual independence fails)} \end{aligned}$$ □

[6] Kolmogorov makes this reference in [6, p.11].

So pairwise independence is a special case of mutual independence. Kolmogorov in fact views the latter as a special case of yet another more general definition, which follows (Definition 9). It has an intuitive content which is best understood if we adhere to his terminology and distinguish between *independence of events* and *independence of experiments*:

Definition 8 (experiment)**.** An *experiment* (equivalently *decomposition, partition*) $\mathfrak{U} = \{A_1, A_2, ..., A_r\}$ with possible results $A_1, ..., A_r$ is a collection of mutually disjoint random events whose union is Ω (ie they are jointly exhaustive).

Definition 9. Consider n experiments $\mathfrak{U}^{(1)}, \mathfrak{U}^{(2)}, ..., \mathfrak{U}^{(n)}$ where:

$$\mathfrak{U}^{(i)} = \{A_1^{(i)}, ..., A_{r_i}^{(i)}\}$$

The $\mathfrak{U}^{(i)}$'s are *(mutually) independent* iff:

$$P(A_{q_1}^{(1)} \cap A_{q_2}^{(2)} \cap ... \cap A_{q_n}^{(n)}) = P(A_{q_1}^{(1)})P(A_{q_2}^{(2)})...P(A_{q_n}^{(n)}) \tag{2.3}$$

for all valid choices of q_i (ie such that $q_1 \leq r_1, ..., q_n \leq r_n$).

Recall that in the case of n mutually independent events $A_1, ..., A_n$, it follows directly that any subcollection of events $A_{i_1}, ..., A_{i_m}$, $m < n$, will also consist of mutually independent events. The same holds for mutual independence of experiments, although it requires a few lines of proof:

Proposition 8. *If $\mathfrak{U}^{(1)}, ..., \mathfrak{U}^{(n)}$ are independent then so are $\mathfrak{U}^{(i_1)}, ..., \mathfrak{U}^{(i_m)}$ for any $m < n$ and distinct i_j's.*

Proof. What needs to be established is this:

$$P(A_{q_1}^{(i_1)} \cap A_{q_2}^{(i_2)} \cap ... \cap A_{q_m}^{(i_m)}) = P(A_{q_1}^{(i_1)})P(A_{q_2}^{(i_2)})...P(A_{q_m}^{(i_m)}) \tag{2.4}$$

We do it by induction on n. The base case is trivial, so we proceed to prove the inductive step. We need only here establish the case where $m = n - 1$ (since if $m < n - 1$, the result follows by the inductive hypothesis on $n - 1$). In this case, we can reorder the i_j's so that $i_1 = 1, ..., i_{n-1} = n - 1$. Then rewriting the LHS of (2.4):

$$\begin{aligned}
P(A_{q_1}^{(1)} \cap ... \cap A_{q_{n-1}}^{(n-1)}) &= \sum_{q=1}^{r_{n-1}} P(A_{q_1}^{(1)}...A_{q_{n-1}}^{(n-1)}A_q^{(n)}), \text{ by total probability} \\
&= \sum_{q=1}^{r_{n-1}} P(A_{q_1}^{(1)})...P(A_{q_{n-1}}^{(n-1)})P(A_q^{(n)}), \text{ by mutual independence} \\
&= P(A_{q_1}^{(1)})...P(A_{q_{n-1}}^{(n-1)}) \sum_{q=1}^{r_{n-1}} P(A_q^{(n)}) \\
&= P(A_{q_1}^{(1)})...P(A_{q_{n-1}}^{(n-1)}), \text{ since } \sum_{q=1}^{r_{n-1}} P(A_q^{(n)}) = P\left(\bigcup_q A_q^{(n)}\right) = 1
\end{aligned}$$

□

We now establish that indeed the notion of mutual independence of events is a special case of that of mutual independence of experiments:

Proposition 9. *Events $A_1, ..., A_n$ are (mutually) independent iff the experiments $\{A_1, A_1^c\}$, ..., $\{A_n, A_n^c\}$ are.*

Proof. The '$\Leftarrow$' direction is given by (2.4). The '$\Rightarrow$' direction is given by an easy inductive argument. We omit the full details and only give the base case. Assume independence of the *events* A_1, A_2, ie: $P(A_1A_2) = P(A_1)P(A_2)$. Then:

$$\begin{aligned} P(A_1 \cap A_2^c) = P(A_1 \backslash (A_1 \cap A_2)) &= P(A_1) - P(A_1 \cap A_2), \text{ by Proposition 1} \\ &= P(A_1) - P(A_1)P(A_2), \text{ by independence of the events} \\ &= P(A_1)(1 - P(A_2)) \\ &= P(A_1)P(A_2^c) \text{ by Proposition 1} \end{aligned}$$

So $P(A_1 \cap A_2^c) = P(A_1)P(A_2^c)$. Similarly we obtain $P(A_1^c \cap A_2) = P(A_1^c)P(A_2)$ and $P(A_1^c \cap A_2^c) = P(A_1^c)P(A_2^c)$. The conjunction of these four statements yields independence of the *experiments*, as required. □

The intuitive content of Kolmogorov's definition and choice of terms is captured in the Proposition below, which combines the definitions of experiments, elementary conditional probability and independence of experiments:

Proposition 10. *A sequence of n experiments are independent iff the probability of the result of each experiment, conditional on the fact that several other experiments have had definite results (each of which has nonzero probability) is equal to the absolute probability, where no conditional assumptions are being made.*

Proof. Formally, this proposition says that, assuming all events in the experiments $\mathfrak{U}^{(1)}, ..., \mathfrak{U}^{(n)}$ have nonzero probability, then the experiments are independent iff for each (experiment) i and each (result) q:

$$P(A_q^{(i)}) = P_{A_{q_1}^{(i_1)} ... A_{q_m}^{(i_m)}}(A_q^{(i)}) \tag{2.5}$$

for each valid choice of $i_1, ..., i_m$ (that does not include i) and each valid choice of $q_1, ..., q_m$. Now (2.5) follows from Proposition 8 by the definition of elementary conditional probability. Conversely, (2.5) yields (2.3) using the Multiplication Theorem. □

Kolmogorov's definition has the advantage of capturing standard intuitions about experiments. In modern textbooks, however, a different definition of independence is used, which can be easily shown to generalise Kolmogorov's notion:

Definition 10. Fix a probability space $(\Omega, \mathfrak{F}, P)$. A countable collection of sub-σ-algebras $\mathfrak{F}_1, \mathfrak{F}_2, ...$ all contained in $\mathfrak{F}$, are said to be *independent*, iff for any $1 \leq i_1 \leq ... \leq i_n$ and sets $G_{i_j} \in \mathfrak{F}_j$, we have:

$$P(G_{i_1} \cap G_{i_2} \cap ... \cap G_{i_n}) = P(G_{i_1})P(G_{i_2})...P(G_{i_n})$$

Proposition 11. *Fix a probability space $(\Omega, \mathfrak{F}, P)$, two*[7] *experiments $\mathfrak{U}^{(1)} = \{A_1, ..., A_n\}$ and $\mathfrak{U}^{(2)} = \{B_1, ..., B_m\}$ and define the following two collections:*

$$\mathfrak{F}_1 =_{df} \{\emptyset, \bigcup_{i \in J} A_i \mid J \subseteq \{1, ..., n\}\}, \quad \mathfrak{F}_2 =_{df} \{\emptyset, \bigcup_{i \in J} B_i \mid J \subseteq \{1, ..., m\}\}$$

Then $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are sub-σ-algebras[8] *of $\mathfrak{F}$. Moreover, $\mathfrak{U}^{(1)}$ and $\mathfrak{U}^{(2)}$ are independent (Definition 9) iff $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are (Definition 10).*

Proof. First we show that $\mathfrak{F}_1$ (and hence also $\mathfrak{F}_2$ by symmetry) is a sub-σ-algebra of $\mathfrak{F}$. Clearly $\mathfrak{F}_1 \subseteq \mathfrak{F}$, by closure of $\mathfrak{F}$. So it suffices to show that $\mathfrak{F}_1$ is a σ-algebra. Since $\bigcup_i A_i = \Omega$ by definition of an experiment, $\Omega \in \mathfrak{F}_1$. As for countable unions of (nonempty) sets in $\mathfrak{F}_1$, by elementary set theory

$$\bigcup_{k \in \mathbb{N}} (\bigcup_{i \in J_k} A_i) = \bigcup_{i \in J} A_i, \quad \text{where } J =_{df} \bigcup_{k \in \mathbb{N}} J_k \subseteq \{1, ..., n\}.$$

So $\mathfrak{F}_1$ is closed under countable unions; for complements we observe that

$$\left(\bigcup_{i \in J} A_i\right)^c = \bigcup_{i \in J^c} A_i, \quad \text{by disjointness of the } A_i\text{'s.}$$

which completes the proof that $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are both sub-σ-algebras of $\mathfrak{F}$. It is immediate that if these are independent (Definition 10), then the experiments $\mathfrak{U}^{(1)}$, $\mathfrak{U}^{(2)}$ also are (Definition 9). For the converse, assume that $A \in \mathfrak{F}_1$, $B \in \mathfrak{F}_2$. If either is empty, trivially $P(A \cap B) = P(A)P(B) = 0$. Otherwise,

$$A =_{df} \bigcup_{i \in I} A_i, \; B =_{df} \bigcup_{i \in J} B_i, \quad \text{for some } I \subseteq \{1, ..., n\}, \; J \subseteq \{1, ..., m\}.$$

Then

$$A \cap B = \bigcup_{i \in I} A_i \cap \bigcup_{i \in J} B_i = \bigcup_{(i,j) \in I \times J} A_i \cap B_j \tag{2.6}$$

But we now observe that if $(i, j) \neq (i', j')$, $(A_i \cap B_j) \cap (A_{i'} \cap B_{j'}) = \emptyset$, since at least one intersection of two mutually disjoint sets is involved. Therefore,

$$\begin{aligned}
P(A \cap B) &= \sum_{(i,j) \in I \times J} P(A_i \cap B_j), \text{ the RHS of (2.6) being a union of disjoint sets} \\
&= \sum_{(i,j) \in I \times J} P(A_i)P(B_j), \text{ by independence of } \mathfrak{U}^{(1)}, \mathfrak{U}^{(2)} \\
&= \left(\sum_{i \in I} P(A_i)\right)\left(\sum_{j \in J} P(B_j)\right), \text{ by simple combinatorics} \\
&= P(\bigcup_{i \in I} A_i)P(\bigcup_{j \in J} B_j), \text{ by disjointness of the } A_i\text{'s and the } B_j\text{'s} \\
&=_{df} P(A)P(B), \quad \text{as required for independence of } \mathfrak{F}_1, \mathfrak{F}_2.
\end{aligned}$$

□

[7] A straightforward induction can generalise this result for any finite number of experiments.

[8] $\mathfrak{F}_i$ is the *least* σ-algebra that contains $\mathfrak{U}^{(i)}$; in modern notation, $\mathfrak{F}_i =_{df} \sigma(\mathfrak{U}^{(i)})$.

From Partitions to Probability Functions

We have so far encountered finite partitions of Ω only, in the form of experiments $\mathfrak{U} = \{A_1, ..., A_r\}$. Arbitrary partitions of Ω are of course defined similarly:

Definition 11 (Partition)**.** A family $\mathfrak{U} \subseteq \mathcal{P}(\Omega)$ is a *partition* of Ω iff the elements of $\mathfrak{U}$ are pairwise disjoint and their union is Ω.

Finite, countable and arbitrary partitions alike are usually given to us via the aid of some indexing set (without any loss of generality):

$$\mathfrak{U} =_{df} \{A_i \mid i \in I\}$$

Under such notation, the partition $\mathfrak{U}$ is essentially represented as a function *from* $\mathfrak{U}$ *to* I, where I is some (arbitrary) indexing set. We may represent it equally well by a function u *from* Ω *to* I as follows:

$$u : \omega \mapsto i, \text{ where } i \text{ is such that } \omega \in A_i \tag{2.7}$$

where clearly u is a function iff $\mathfrak{U}$ is a partition. Conversely, given an arbitrary probability function u we can write:

$$\mathfrak{U} =_{df} \{u^{-1}(i) \mid i \in I\} \tag{2.8}$$

where

$$u^{-1}(a) =_{df} \{\omega \in \Omega \mid u(\omega) = a\}$$

Naturally we will also wish to assume that $u^{-1}(i)$ be in $\mathfrak{F}$ for all $i \in I$, since we wish our partition to contain sets in $\mathfrak{F}$ only. This motivates the following definition.

Definition 12 (Probability Function)**.** Let $(\Omega, \mathfrak{F}, P)$ be a probability space. We call a function $u : \Omega \to I$ a **probability function** iff:

$$\text{for all } a \in u[\Omega],\ u^{-1}(a) \in \mathfrak{F}$$

Representing a partition as a function commits oneself to a particular choice of index, *label* so to speak, for each set in the partition. This is not so in the representation of a partition as a set of sets, since two partitions employing different indexing schemes will be equal if they contain precisely the same sets. Such 'relabellings' essentially are bijections between indexing sets, and apart from them, partitions are uniquely represented by functions:

Proposition 12. *Let u_1 and u_2 be two probability functions on Ω. Then the generated partitions $\mathfrak{U}_1$, $\mathfrak{U}_2$ are equal iff there exists a bijective mapping $f : \ u_1[\Omega] \to u_2[\Omega]$ such that $u_2 = f \circ u_1$.*

Proof. This proof is tedious definition-chasing, so we omit it. □

The real advantage of working in terms of probability functions rather than partitions is that it allows us to better exploit any structure that our indexing set may have, such as an order relation and field operations, whereas a partition, as we have explained, understands its indexing set merely as a set of labels. This advantage will become apparent when we introduce in later chapters a special case of highly structure-preserving functions $\Omega \to \mathbb{R}$, *random variables*.

Probability Functions and Measurability: a discussion

We will now prove certain basic results about probability functions, preparing the ground for more advanced work in the next sections. For the purpose of this subsection, we consider throughout two sets of elementary outcomes Ω, Ω' and a function $u : \Omega \to \Omega'$, where we might as well constrain Ω' so that $u[\Omega] = \Omega'$.

Definition 13 (Pre-images)**.** We denote by $u^{-1}[A']$ the *pre-image* or *inverse image* of A' under u:

$$u^{-1}[A'] =_{df} \{\xi \in \Omega \mid u(\xi) \in A'\}$$

We simplify our notation for pre-images of singletons as in the previous section:

$$u^{-1}(a) =_{df} u^{-1}[\{a\}] =_{df} \{\omega \in \Omega \mid u(\omega) = a\}$$

Let us explain briefly why the pre-image operator is the key to this entire discussion. In terms of partitions, we would like to assign probabilities to sets of labels according to the probabilities of the sets in the partition that the labels code for. In formal terms, we assign to each subset of Ω' the probability of its pre-image.

Clearly we can do that precisely for those subsets of Ω' whose pre-images live in our given space $(\mathfrak{F}, P)$. We denote the class of such sets by $\mathfrak{F}^{(u)}$:

$$\mathfrak{F}^{(u)} =_{df} \{A' \subseteq \Omega' \mid u^{-1}[A'] \in \mathfrak{F}\} \tag{2.9}$$

We can now proceed to define on it the following assignment of probabilities:

$$\forall A' \in \mathfrak{F}^{(u)},\ P^{(u)}(A') =_{df} P(u^{-1}[A']) \tag{2.10}$$

It turns out that these definitions in fact suffice to make $(\Omega', \mathfrak{F}^{(u)}, P^{(u)})$ a probability space, as the next two propositions establish.

Proposition 13. *If $\mathfrak{F}$ is a σ-algebra over Ω, then $\mathfrak{F}^{(u)}$ is a σ-algebra over Ω'.*

Proof. We need to show that $\mathfrak{F}^{(u)}$ is closed under complement and countable intersections and unions. It suffices to observe that, by elementary set theory, the pre-image construct commutes with all these operations. Hence, $\mathfrak{F}^{(u)}$ automatically inherits closure from $\mathfrak{F}$. □

Proposition 14. *If $(\Omega, \mathfrak{F}, P)$ is a probability space, then so is $(\Omega', \mathfrak{F}^{(u)}, P^{(u)})$, as defined by* (2.9) *and* (2.10)*.*

Proof. Axiom I holds of $\mathfrak{F}^{(u)}$ by the previous proposition and Axiom III holds of $P^{(u)}$ by construction. Clearly $u^{-1}[\Omega'] = \Omega$ so $\Omega' \in \mathfrak{F}^{(u)}$ and $P^{(u)}(\Omega') = 1$, which yields Axioms II and IV.

Finally, we show countable additivity (which yields Axioms V and VI). For any countable collection $A_1, A_2, ...$ such that the A_i's are pairwise disjoint,

$$u^{-1}\left(\bigcup_i A_i\right) = \bigcup_i u^{-1}[A_i] \ \text{ where the } u^{-1}[A_i]\text{'s are pairwise disjoint,} \tag{2.11}$$

since the pre-image not only commutes with unions but also with intersections, hence preserving disjointness. Then

$$\begin{aligned} P^{(u)}(\bigcup_i A_i) &=_{df} P(u^{-1}[\bigcup_i A_i]) \\ &= P(\bigcup_i u^{-1}[A_i]), \text{ since pre-images commute with disjoint sums} \\ &= \sum_i P(u^{-1}[A_i]), \text{ by countable additivity of } P \text{ and (2.11)} \\ &=_{df} \sum_i P^{(u)}(A_i). \end{aligned}$$

□

Given a probability space over the domain of u, we have managed to induce a probability space over its image, i.e.,

$$P^{(u)}(A') =_{df} P(u(\omega) \in A') =_{df} P(u^{-1}[A'])$$

Remark. Observe that the leftmost and rightmost expressions both represent probabilities of *sets*, whereas the middle expression is the probability of the *proposition* "$u(\omega) \in A'$". It will often be the case, as is with "$u(\xi) \in A'$", that a certain proposition S will correspond in a definite manner to the statement "$\xi \in B$" where $B \in \mathfrak{F}$ for a certain space $(\Omega, \mathfrak{F}, P)$. In such cases, we shall denote the probability of that proposition by $P(S)$. In fact, it is only such propositions that are allowed to have probabilities at all according to the Grundbegriffe[9].

Having finalised our construction of the induced space $(u[\Omega], \mathfrak{F}^{(u)}, P^{(u)})$, we now prove a couple of technical results to familiarise the reader with the properties of this construction. Firstly, everything works well with compositions:

Proposition 15. *Let $u_1 : \Omega \to \Omega'$, $u_2 : \Omega' \to \Omega''$ and consider:*

$$u =_{df} u_2 \circ u_1 : \Omega \to \Omega''$$

Then the following holds true:

$$P^{(u)}(A'') = P^{(u_1)}(u_2^{-1}(A'')), \qquad (A'' \subseteq \Omega'')$$

Proof. Definition-chasing:

$$u(x) \in A'' \Leftrightarrow u_2(u_1(x)) \in A'' \Leftrightarrow u_1(x) \in u_2^{-1}(A'') \Leftrightarrow x \in u_1^{-1}(u_2^{-1}(A'')) \quad \square$$

It is also important to remark that the construction $\mathfrak{F}^{(u)}$ equals the whole of the power set of $u[\Omega]$ whenever u is a probability function and $u[\Omega]$ is countable:

Proposition 16. *Let $\mathfrak{F}$ be a σ-algebra over Ω and $u : \Omega \to u[\Omega]$ be a probability function. If $u[\Omega]$ is countable, then $\mathfrak{F}^{(u)} = \mathcal{P}(u[\Omega])$.*

[9] This is in contrast to the Bayesian doctrine, which has entirely different and much weaker criteria for the admissibility of an assignment of probabilities on propositions.

Proof. By definition of a probability function, $\mathfrak{F}^{(u)}$ contains all singletons of $u[\Omega]$. Since $u[\Omega]$ is countable, it must also contain all subsets of $u[\Omega]$, by countable closure of $\mathfrak{F}^{(u)}$. □

Whenever $u[\Omega]$ is not countable, $\mathfrak{F}^{(u)}$ will in general be much smaller than $\mathcal{P}(u[\Omega])$. A common error, for instance, is to assume that it must still contain all *images* of elements in $\mathfrak{F}$. An interesting but nontrivial counterexample to this statement for $u[\Omega] = \mathbb{R}$ can be found in Proposition 46, in the Appendix.

A Remark on Measurability

The notion of a function from one probability space to another underpins much of probability theory. The key factor is the way in which the pre-image of a certain function u and the σ-algebras on either space relate. Kolmogorov's approach to the study of this problem is the one we have taken so far, whereby:

Definition 14. Fix a space $(\Omega, \mathfrak{F}, P)$ and a probability function $u : \Omega \to \Omega'$. Then the *induced σ-algebra by u* is denoted by $\mathfrak{F}^{(u)}$ and give by:

$$\mathfrak{F}^{(u)} =_{df} \{A \mid u^{-1}[A] \in \mathfrak{F}\}$$

This is not the only possible set-up. In some contexts, it is convenient instead to fix a σ-algebra $\mathfrak{F}'$ over the image space and construct a σ-algebra over the domain large enough to contain all pre-images under u:

Definition 15. Fix $(\Omega', \mathfrak{F}')$ and $u : \Omega \to \Omega'$. Then define:

$$\sigma(u) =_{df} \{u^{-1}[A] \mid A \in \mathfrak{F}'\}$$

This is a σ-algebra since the pre-image commutes with countable set operations.

Either of these two approaches serves to induce a new space from some given space[10]. In modern probability, however, the focus has shifted away from the construction of new spaces and now lies with the study of well-behaved functions between pairs of standardised spaces. One then fixes two probability spaces and is invited to consider functions from one to the other that are *measurable*:

Definition 16. Fix two σ-algebras $\mathfrak{F}$ and $\mathfrak{F}'$ over two sets Ω and Ω' respectively. A function $u : \Omega \to \Omega'$ is said to be $\mathfrak{F}'/\mathfrak{F}$-*measurable* or simply *measurable* iff:

$$\forall A' \in \mathfrak{F}',\ u^{-1}[A'] \in \mathfrak{F}$$

Presently we show that Definitions 14, 15 and 16 can translate to one another, an indication that they are but alternative approaches to the same study.

[10] Observe however that the construction of $\sigma(u)$ does not automatically equip us with a probability measure over Ω, induced by u, as is the case in the construction of $\mathfrak{F}^{(u)}$. It usually takes a lot more work to prove the existence of such a measure, since $\sigma(u)$ might be too large (as large as the power set of Ω, see the proof of Proposition 17) to admit a measure.

Proposition 17. *Fix two σ-algebras $\mathfrak{F}$ and $\mathfrak{F}'$ over two sets Ω and Ω' respectively. Then, for any function $u : \Omega \to \Omega'$:*

(a) the σ-algebra $\mathfrak{F}^{(u)}$ is maximal such that u may be $\mathfrak{F}'/\mathfrak{F}$-measurable,

(b) the σ-algebra $\sigma(u)$ is least such that u may be u $\mathfrak{F}'/\mathfrak{F}$-measurable,

(c) u is $\mathfrak{F}'/\mathfrak{F}$-measurable iff $\mathfrak{F}' \subseteq \mathfrak{F}^{(u)}$.

(d) u is $\mathfrak{F}'/\mathfrak{F}$-mesaurable iff $\sigma(u) \subseteq \mathfrak{F}$.

Proof. We need only express $\mathfrak{F}^{(u)}$ and $\sigma(u)$ as extrema of families of σ-algebras:

$$\mathfrak{F}^{(u)} = \bigcup \mathcal{G}, \quad \text{where } \mathcal{G} =_{df} \{\Sigma \mid u \text{ is } \Sigma/\mathfrak{F}\text{-measurable}\}$$

$$\sigma(u) = \bigcap \mathcal{H}, \quad \text{where } \mathcal{H} =_{df} \{\Sigma \mid u \text{ is } \mathfrak{F}'/\Sigma\text{-measurable}\}$$

where neither family can be empty, since trivially $\{\emptyset, \Omega\} \in \mathcal{G}$ and $\mathcal{P}(\Omega) \in \mathcal{H}$. □

Remark. We can naturally generalise the Definition of $\sigma(u)$ for the case where we are given a family of probability functions which we want to make measurable:

Proposition 18. *Fix $(\Omega, \mathfrak{F})$ and a family of functions $u_\gamma : \Omega \to \Omega_\gamma$ indexed by $\gamma \in \Gamma$, each image associated with a σ-algebra $\mathfrak{F}_\gamma$. Then define:*

$$\sigma(u_\gamma : \gamma \in \Gamma) =_{df} \bigcap \{\mathfrak{F} \mid u \text{ is } \mathfrak{F}_\gamma/\mathfrak{F}\text{-measurable for all } \gamma \in \Gamma\}$$

This is the least σ-algebra that makes all of u_γ measurable.

Proof. That $\sigma(u_\gamma : \gamma \in \Gamma)$ is a σ-algebra follows from the trivial observation that the intersection of a non-empty arbitrary collection of σ-algebras is a σ-algebra. That the collection is non-empty is guaranteed by the fact that the power set $\mathcal{P}(\Omega)$ is trivially a σ-algebra that makes any function measurable. □

2.2 Constructing Probability Spaces

In this section we will study methods and theorems that can assist us in the constructions of probability spaces. These results will prove essential in the following section, wherein we will be investigating concrete examples.

The Caratheodory Extension Theorem: a discussion

It is usually much easier mathematically to fully specify a probability assignment on an algebra, than it is on a σ-algebra (recall Definitions 2 and 3). In fact, σ-algebras themselves are rarely if ever given explicitly, but rather as the countable closure of some smaller family of subsets of Ω (often an algebra):

Proposition 19. *For any family S of subsets of Ω, the following family denoted by $\sigma(S)$ is the least σ-algebra containing S:*

$$\sigma(S) =_{df} \bigcap \mathcal{F}, \quad \textit{where } \mathcal{F} =_{df} \{\mathfrak{F} \mid \mathfrak{F} \textit{ is a } \sigma\textit{-algebra over } \Omega \textit{ and } S \in \mathfrak{F}\}$$

We call S the generator *of $\sigma(S)$ and $\sigma(S)$ the* closure[11] *of S.*

Remark. The notation $\sigma(S)$ can always be distinguished from the notation $\sigma(u)$ (Definition 15) from the fact that S is a set and u a function. No clearer standard notational distinction exists.

Proof. Firstly note that $\mathcal{F} \neq \emptyset$, since $\mathcal{F}$ trivially contains the power set $\mathcal{P}(\Omega)$. We now need only show that $\sigma(S)$ is a σ-algebra over Ω, since then it is direct from definition that it is least. Clearly $\Omega \in \sigma(S)$, because Ω can be found in every $\mathfrak{F} \in \mathcal{F}$. Moreover, any collection $A_1, A_2, ... \in \sigma(S)$ can also be found in each $\mathfrak{F} \in \mathcal{F}$. Hence, by closure of all such $\mathfrak{F}$, also the result of any countable operation on the collection $A_1, A_2, ...$ must be in each $\mathfrak{F}$ and hence also in S. □

Our hope is that we can define probability measures over relatively simple generators, which would then uniquely and consistently extend to probability measures over their closures via the properties of measure. Before we establish how that can be done, we ensure that we have a proper understanding of what it means for P to be a probability measure over S, for choices of S that are not σ-algebras:

Definition 17. Let S be any family of subsets of Ω containing Ω and $\emptyset$. Then P is a probability measure over S if $P(\Omega) = 1 - P(\emptyset) = 1$ and P is *countably additive* over S in the following sense:

$$\text{if } A_1, A_2, ... \in S \textit{ and } \bigcup_i A_i \in S, \quad \text{then } P(\bigcup_i A_i) = \sum_i P(A_i)$$

Note that $\bigcup_i A_i \in S$ now forms a part of the assumption.

On the one hand, it is clear enough that any assignment of probabilities on S will uniquely determine certain probabilities in $\sigma(S)$, via the properties of measure. On the other, such forced assignments also raise possibilities of

[11] The definition of $\sigma(S)$ as an infimum over a collection of sets is probably more recent than Kolmogorov's Grundbegriffe. At that time, mathematicians were more likely to use the equivalent 'constructive' definition which uses transfinite induction:

Theorem (Inductive Definition of Countable Closure)**.** *Given S, a family of subsets of Ω we can construct $\sigma(S)$ setting $\mathfrak{F}_1 = S$ and using the following recursion:*

$$\textit{for } \xi \textit{ a successor ordinal, } \mathfrak{F}_{\xi+1} =_{df} \{\textit{countable unions of complements of sets in } \mathfrak{F}_\xi\}$$

$$\textit{for } \lambda \textit{ a limit ordinal, } \mathfrak{F}_\lambda =_{df} \{\textit{countable unions of sets in } \bigcup_{\xi<\lambda} \mathfrak{F}_\xi\}$$

$$\sigma(S) =_{df} \bigcup_\xi \mathfrak{F}_\xi$$

conflict, since there might be more than one ways to produce a certain set A in $\sigma(S)$ from sets in S, each yielding a different probability for A.

We therefore require two results: firstly, that no two forced assignments contradict each other (existence of an extension); Secondly, that the probabilities of all sets in $\sigma(S)$ be forced (uniqueness of the extension).

When the generator is a σ-algebra both uniqueness and existence are trivially satisfied. We are hence naturally led to ask: how much, if at all, weaker can the structure of the generator S be, so that we are guaranteed uniqueness and existence of an extension over $\sigma(S)$?

In the three results that follow, we separately answer this question for the case of uniqueness and the case for existence, since it turns out that the latter is a strictly stronger requirement[12]: existence requires finite closure, whereas uniqueness only requires closure under finite intersections.

Definition 18 (π-system)**.** A collection $\mathcal{I}$ of subsets of Ω is a π-system over Ω whenever it is closed under pairwise intersections.

Theorem 1 (Uniqueness Lemma)**.** *Let $\mathcal{I}$ be a π-system over Ω and let P, P' be two probability measures on $\sigma(\mathcal{I})$. Then:*

$$P = P' \text{ on } \mathcal{I} \;\Leftrightarrow\; P = P' \text{ on } \sigma(\mathcal{I})$$

Proof. Omitted, can be found in [15, p.194]. □

Proposition 20. *There exists a π-system $\mathcal{I}$ and a probability measure P over $\mathcal{I}$ such that no extension of P to a probability measure over $\sigma(\mathcal{I})$ exists.*

Proof. Let $\Omega = \{\omega_1, \omega_2, ...\}$ and $\mathcal{I} = \{\Omega, \emptyset, \{\omega_1\}, \{\omega_2\}, ...\}$. This is a π-system. Now consider the following assignment P of probabilities on $\mathcal{I}$:

$$P(\Omega) =_{df} 1 - P(\emptyset) =_{df} 1, \quad \text{and for all } i,\ P(\omega_i) =_{df} 0$$

This assignment agrees with Definition 17, because any family of sets in $\mathcal{I}$ whose union remains in $\mathcal{I}$ must either contain only one non-empty set or it must contain Ω; either way, countable additivity is satisfied:

$$P(A \cup \emptyset) = P(A) = P(A) + P(\emptyset), \quad \text{for any } A \in \mathcal{I}$$

$$P(\Omega \cup \{\omega\} \cup ... \cup \{\omega'\}) = P(\Omega) = P(\Omega) + P(\{\omega\}) + ... + P(\{\omega'\})$$

However, there exists no probability measure P' over $\sigma(\mathcal{I}) = \mathcal{P}(\Omega)$ that extends P, since if that were so, from countable additivity of P' we should get:

$$1 = P'(\Omega) = P(\bigcup_i \{\omega_i\}) = \sum_i P'(\{\omega_i\}) = \sum_i P(\{\omega_i\}) = 0$$ □

Theorem 2 (Caratheodory Extension)**.** *Let Σ be an algebra over Ω and P be a probability measure on Σ. Then there exists a (unique) probability measure on $\sigma(\Sigma)$ that agrees with P on Σ.*

Proof. Proof omitted, can be found in [15, pp.195-199]. Uniqueness follows by Uniqueness Lemma, since an algebra is trivially also a π-system. □

[12] This makes sense, since we need S to have enough structure to reveal any conflicting information that might be implicit in P.

Remark on Kolmogorov's Version of Definition 1

Equipped with the Carathéodory Extension Theorem, we are now in a position to compare Kolmogorov's approach to the axiomatic definition of probability spaces with the modern approach (Definition 1).

The results we have just established heavily rely on the presence of countable additivity (or equivalently Axiom VI). However, at the time of the writing of the Grundbegriffe, countable additivity was still generally being viewed as a separate hypothesis, useful for probabilistic reasoning in pure mathematics but otherwise lacking empirical content[13]. This explains the reason why Kolmogorov in 1933 goes about the definition of probability spaces in a roundabout way:

A. Initially he works with finite closure only, defining what he called *Generalised Probability Fields*, ie probability spaces over *algebras* that do not necessarily satisfy Axiom VI.

B. He then introduces Axiom (VI) as an extra "arbitrary" hypothesis, "found expedient in researches of the most diverse sort"[14]. This produces the definition of *Probability Fields*.

C. Finally, he arbitrarily restricts[15] his attention to probability fields over σ-algebras, which he called *Borel Probability Fields*.

It is only this latter notion of a 'Borel Probability Field' that agrees with our notion of a Probability Space. Moreover, the extension theorem serves precisely to establish that any 'Probability Field' uniquely corresponds to a 'Borel Probability Field', although the same cannot be said for an arbitrary 'Generalised Probability Field', since there we lack Axiom VI and the Extension Theorem does not go through. In fact, our formalism has no place for 'Generalised Probability Fields'.

[13] Certain philosophers of probability have argued that since only finitely complex composite events are ever observed, the full strength of countable closure makes it much harder to empirically justify a choice of probability assignment P, since there will typically be an infinity of specific probability values which we will be even in principle unable to empirically verify. Axiom (VI) was similarly viewed as a philosophical compromise. Kolmogorov describes the whole matter very succintly as follows:

> Since the new axiom is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning. [...] For, in describing any observable random process we can obtain only finite fields of probability. [...] *We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI.* This limitation has been found expedient in researches of the most diverse sort.

[14] See previous footnote.

[15] In fact, he views this not as a restriction, but as an *extension*, claiming that sets of events that form algebras may admit of an empirical interpretation, but σ-algebras are too large to do so, since they invoke infinity in an essential way. This may seem odd to the modern reader since σ-algebras are formally *special cases* of algebras, and better behaved, if anything. However, with a 'constructive' definition of algebras in mind, as described in footnote 11 earlier, a recursively defined algebra only requires finite induction, whereas a σ-algebra requires transfinite induction. In this context, an algebra is indeed a simpler object, which might explain Kolmogorov's reluctance to deal with σ-algebras straight from the onset.

Remark. The term 'Borel' is at present used only in topological contexts, as the reader will be reminded soon. For this reason we abandoned Kolmogorov's original terminology for its modern equivalent, 'Probability Space'.

Spaces with a Countable set of Elementary Outcomes

Let $\Omega = \{\xi_1, \xi_2, ..., \}$ and p_1, p_2, ... be a sequence of positive real numbers such that:

$$\sum_n p_n = 1 \tag{2.12}$$

Then, assuming that we wish our σ-algebra $\mathfrak{F}$ to contain all singletons, the following assignment of probabilities on the singletons can be uniquely extended to a probability measure over $\mathfrak{F}$:

$$\forall\, n \in \mathbb{N},\ P(\{\xi_n\}) =_{df} p_n \tag{2.13}$$

Proposition 21. *There exists precisely one probability space $(\Omega, \mathfrak{F}, P)$ such that $\mathfrak{F}$ contains the singletons and P satisfies (2.13).*

Proof. Firstly, the closure of the set of singletons is the power set $\mathcal{P}(\Omega)$, which is also the maximal σ-algebra over Ω, so it is the only possible choice for $\mathfrak{F}$. Now for any set S in $\mathfrak{F}$, we define the following probability:

$$P(S) =_{df} \sum_{\xi \in S} P(\{\xi\})$$

Observe that the infinite sum on the RHS converges absolutely always, since it only consists of positive terms and the sequence of partial sums is monotonically increasing and, by (2.12), bounded above by 1.

In fact, P is a probability measure. By (2.12), $P(\Omega) = 1$ and, by definition, $P(\emptyset) = 0$. Moreover, if $(A_i : i \in \mathbb{N})$ is a sequence of disjoint sets in $\mathfrak{F}$, then:

$$P(\bigcup_i A_i) =_{df} \sum_{\xi \in \bigcup_i A_i} P(\{\xi\}) = \sum_i \sum_{\xi \in A_i} P(\{\xi\}) =_{df} \sum_i P(A_i)$$

by absolute convergence and disjointness of the A_i's. This completes the proof that P is a probability measure. Finally, clearly P agrees with (2.13) on all singletons, which together with the empty set form a π-system, so it is the unique measure that satisfies (2.13), as required. □

Spaces with $\mathbb{R}$ as a set of Elementary Outcomes

The standard σ-algebra to work with over the reals and by far the most important example of a σ-algebra in this dissertation is the following:

$$\sigma(\text{open sets in } \mathbb{R})$$

This is in fact a special case of a general construction that works for any *topological space*. We recall the definition of a topological space:

Definition 19. A *topological space* $\{\Omega, \mathcal{T}\}$ consists of a non-empty set Ω together with a fixed collection $\mathcal{T}$ of subsets of A satisfying:

T1. $\Omega, \emptyset \in \mathcal{T}$

T2. $\mathcal{T}$ is closed under finite intersections.

T3. $\mathcal{T}$ is closed under arbitrary unions.

We say a set $A \subseteq \Omega$ is *open* iff $A \in \mathcal{T}$.

Much of the theory of spaces such as $\mathbb{R}$ that are naturally endowed with a topology rests on being able to reason with open sets. It is hence natural whenever Ω can be seen as a topological space that we should consider σ-algebras that contain the topology of Ω. The least such σ-algebra is called *Borel*:

Definition 20. Let $(\Omega, \mathcal{T})$ be a topological space. Then the *Borel σ-algebra* over Ω is defined as follows:

$$\mathcal{B}(\Omega) =_{df} \sigma(\mathcal{T})$$

We call any set $A \subseteq \Omega$ that is in the collection $\mathcal{B}(\Omega)$ a *Borel set of* Ω.

It will prove useful that in the special case where $\Omega = \mathbb{R}$, the Borel sets can be generated by the set of **half-rays**, $\{(-\infty, a) \mid a \in \mathbb{R}\}$:

Proposition 22.

$$\mathcal{B}(\mathbb{R}) =_{df} \sigma(\{\textit{open sets in } \mathbb{R}\}) = \sigma\left(\{(-\infty, a) \mid a \in \mathbb{R}\}\right) \tag{2.14}$$

Proof. This proof relies on the simple remark that for any two sets G and H:

$$\text{if } G \subseteq \sigma(H) \text{ and } H \subseteq \sigma(G), \quad \text{then } \sigma(G) = \sigma(H)$$

We will be using this in several proofs. In this case, we need to establish that:

$$\{\text{half-rays in } \mathbb{R}\} \subseteq \sigma(\{\text{open sets in } \mathbb{R}\}) \tag{A}$$
$$\{\text{open sets in } \mathbb{R}\} \subseteq \sigma(\{\text{half-rays in } \mathbb{R}\}) \tag{B}$$

In one direction, (A) follows trivially since half-rays are open sets. In the other, it suffices to establish that all open intervals (a, b) can be generated from half-rays, since any open set can be expressed as a countable union of open intervals. However, we readily observe that:

$$(a, b) = (-\infty, b) \setminus (-\infty, a]$$

whereby the following completes the proof:

$$(-\infty, a] = \bigcap_n (-\infty, a + \frac{1}{n}) \qquad \square$$

Remark. By a similar proof one may show that $\mathcal{B}(\mathbb{R}) = \sigma(\{(-\infty, a] \mid a \in \mathbb{R}\})$.

The foundation for the construction of probability measures on $\mathbb{R}$ is, naturally, the notion of *length*, as generalised into a countably additive set function by Lebesgue. To obtain a *probability* measure out of the notion of length we have of course to limit ourselves to a bounded interval. We hence prove the existence of the *Lebesgue probability measure* on $[0, 1]$, with obvious generalisations to any bounded interval (a, b), $[a, b)$, $(a, b]$ or $[a, b]$.

Theorem 3 (Existence of Lebesgue Measure)**.** *There exists a unique probability measure Leb on* $([0, 1], \mathcal{B}([0, 1]))$ *such that, for all* $0 \leq a \leq 1$*, we have:*

$$Leb([0, a)) = a \tag{2.15}$$

Proof. The proof is nontrivial and standard so we will not include it in full in the dissertation - we offer precisely enough information for the reader to be able to understand what the nontrivial part is.

Let $\Omega = [0, 1]$. We will construct an *algebra* Σ_0 over Ω as follows. We let $A \in \Sigma_0$ iff A may be written as a finite disjoint union of open intervals and singletons:

$$A = (a_1, a_2) \cup \{a_3\} \cup (a_4, a_5) \cup ... \cup (a_{n-1}, a_n) \tag{A}$$

where $0 \leq a_i \leq a_{i+1} \leq 1$. It is routine to establish that Σ_0 is in fact an algebra (i.e., finitely closed). We then define the following assignment Leb on open intervals and points:

$$Leb((a, b)) =_{df} b - a, \qquad Leb(\{a\}) = 0$$

This extends to an assignment of probabilities to Σ_0 simply via finite addition. It is then easy to establish that this assignment $Leb(A)$ for $A \in \Sigma_0$ is stable under the various ways in which A can be represented as a disjoint union of points and open intervals[16]. Having thus established that P is well-defined, we trivially observe that:

$$Leb(\Omega) = 1,\ Leb(\emptyset) = 0 \text{ and } Leb \text{ is finitely additive}$$

Presently the nontrivial part of this proof is required:

Lemma. *Leb is countably additive over* Σ_0*.*

Proof of Lemma. Omitted, can be found in [15, pp.200-202]. This proof makes essential use of the topological properties (namely, the compactness) of $[0, 1]$. □

It then follows from the Carathéodory Extension Theorem that Leb can be extended to $\sigma(\Sigma_0)$, which is of course precisely $\mathcal{B}([0, 1])$ by Proposition 22. Finally, it follows from the Uniqueness Lemma that Leb thus defined is unique, since (2.15) determines Leb on the π-system $\{[0, a) \mid 0 \leq a \leq 1\}$. □

In later sections, we will make essential use of Lebesgue measure to derive a general method for the construction of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, by way of their correspondence with *distribution functions*.

[16] One may argue on the basis of the observation that overlapping endpoints cancel out.

Remark on Completions

Note that the space we have just constructed, $([0,1], \mathcal{B}([0,1]), Leb)$, is *not* a complete space in the measure-theoretic sense (see Appendix A) and hence $\mathcal{B}([0,1])$ is strictly *smaller* than what is known as the family of *Lebesgue measurable sets*. Naturally the complete space $([0,1], Leb([0,1]), Leb)$ is also a probability space. However, probability theorists tend to prefer to avoid completions unless they are needed for the mathematics. In particular, for topological spaces, the Borel σ-algebra is sufficient for all important theorems and completions are mostly considered an unnecessary complication that results only in loss of tangibility.

Product and Component Spaces

Joint consideration of two sets of elementary outcomes Ω_1, Ω_2 can be achieved by encoding our experiment with pairs of values:

$$(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$$

Naturally, this same representation also gives an 'encoding' of events involving, say, Ω_1 only, obtained by disregarding second coordinates:

$$Pr(X_1) =_{df} P(\{(x', y) \in \Omega_1 \times \Omega_2 \mid x' \in X_1\}) = P(X_1 \times \Omega_2), \quad (X_1 \subseteq \Omega_1)$$

Formally, this encoding is the pre-image $\pi_1^{-1}(X_1)$ of X_1 under the *projection* π_1. This invites us to consider the projection maps as probability functions:

$$\pi_1 : (\omega_1, \omega_2) \mapsto \omega_1, \quad \pi_2 : (\omega_1, \omega_2) \mapsto \omega_2$$

On the basis of this observation, we proceeed to investigate the relationship between distinct spaces over Ω_1 and Ω_2 and spaces 'joint' over $\Omega_1 \times \Omega_2$.

Component Spaces

Assume we are given a fixed probability space $(\Omega_1 \times \Omega_2, \mathfrak{F}, P)$. It is easy to construct its *component spaces* over each of Ω_1 and Ω_2. As we observed, the encodings in $\mathfrak{F}$ of sets in $\mathfrak{F}_i$ are just pre-images under the projection maps. In accordance with our results on probability functions, we can then see that the two probability spaces we are looking for are:

$$(\Omega_1, \mathfrak{F}^{(\pi_1)}, P^{(\pi_1)}) \quad \text{and} \quad (\Omega_2, \mathfrak{F}^{(\pi_2)}, P^{(\pi_2)}) \tag{2.16}$$

Product Spaces

Fix now two probability spaces $(\Omega_1, \mathfrak{F}_1, P_1)$, $(\Omega_2, \mathfrak{F}_2, P_2)$. We wish to define a σ-algebra over $\Omega_1 \times \Omega_2$ that contains 'encodings' of all sets A_i in $\mathfrak{F}_i$ - formally, their pre-images under the respective projections. The least such algebra is precisely the object defined in Definition 15:

Definition 21. Let $\mathfrak{F}_1$, $\mathfrak{F}_2$ be two given σ-algebras over Ω_1, Ω_2 respectively. We define the *product σ-algebra* of $\mathfrak{F}_1$ and $\mathfrak{F}_2$, denoted by $\mathfrak{F}_1 \times \mathfrak{F}_2$, as follows:

$$\mathfrak{F}_1 \times \mathfrak{F}_2 =_{df} \sigma(\pi_1, \pi_2) =_{df} \sigma(\{\pi_i^{-1}(A_i) \mid A_i \in \mathfrak{F}_i\})$$

Proposition 23. *The product σ-algebra can also be given as follows:*

$$\mathfrak{F}_1 \times \mathfrak{F}_2 = \sigma(\{A_1 \times A_2 \mid A_i \in \mathfrak{F}_i\}) \tag{2.17}$$

Proof. As we have observed, the pre-image under projection π_i of a set A_i in $\mathfrak{F}_i$ can be written as a cartesian product $\Omega_1 \times A_2$ or $A_1 \times \Omega_1$ respectively. Conversely, the cartesian product $A_1 \times A_2$ can be written as an *intersection* of pre-images under projections, $\pi_1^{-1}(A_1) \cap \pi_2^{-1}(A_2)$. Therefore:

$$\sigma(\pi_1, \pi_2) =_{df} \sigma(\{\pi_i^{-1}(A_i) \mid A_i \in \mathfrak{F}_i\}) = \sigma(\{A_1 \times A_2 \mid A_i \in \mathfrak{F}_i\}) \qquad \square$$

Naturally, we are now only interested in probability measures P over $\mathfrak{F}_1 \times \mathfrak{F}_2$ that *agree* with P_i on encodings of sets A_i in $\mathfrak{F}_i$. Formally:

$$P(\pi_i^{-1}(A_i)) = P_i(A_i), \qquad (A_i \in \mathfrak{F}_i) \tag{2.18}$$

Of particular importance is the measure that additionally demands 'independence' of the two underlying spaces, known as *product measure*:

Theorem 4 (Existence of Product Measure)**.** *Let $(\Omega_1, \mathfrak{F}_1, P_1)$, $(\Omega_2, \mathfrak{F}_2, P_2)$ be two probability spaces. Then there exists a* unique *product measure P on the product σ-algebra $\mathfrak{F}_1 \times \mathfrak{F}_2$ that satisfies, for all $A_i \in \mathfrak{F}_i$:*

$$P(A_1 \times A_2) = P_1(A_1)P_2(A_2) \tag{2.19}$$

This measure is sometimes denoted by $P_1 \times P_2$.

Proof. Existence follows from Fubini's theorem, which we will state much later (without proof), when we are equipped with the notion of Lebesgue integration. Uniqueness follows from the Uniqueness Lemma, since (2.19) constrains the values of P everywhere on the π-system $\mathcal{I}$ defined below:

$$\mathcal{I} =_{df} \{A_1 \times A_2 \mid A_i \in \mathfrak{F}_i\}$$

and, from Proposition 23, we know that $\mathfrak{F}_1 \times \mathfrak{F}_2 = \sigma(\mathcal{I})$. $\square$

Remark. Observe that among measures that satisfy (2.18), the additional constraint (2.19) can be written as:

$$P(\pi_1^{-1}(A_1) \cap \pi_2^{-1}(A_2)) = P(\pi_1^{-1}(A_1))P(\pi_2^{-1}(A_2))$$

which expresses a requirement of independence of encodings of events $A_1 \in \mathfrak{F}_1$ from events in $A_2 \in \mathfrak{F}_2$ in exactly the sense of Section 2.1.

The product measure is of course not the only measure over $\mathfrak{F}_1 \times \mathfrak{F}_2$ that satisfies (2.18). There usually exists a wide range of such measures for any given choice of spaces $(\Omega_i, \mathfrak{F}_i, P_i)$, reflecting the existence of *correlations* between outcomes of Ω_1 and Ω_2. We will see such examples later in the main text.

An Aside

According to the above, we can say that two spaces $(\Omega_1, P_1, \mathfrak{F}_1)$ and $(\Omega_2, \mathfrak{F}_2, P_2)$ *agree* with a space $(\Omega_1 \times \Omega_2, \mathfrak{F}, P)$ precisely if all encodings of sets $A_i \in \mathfrak{F}_i$ are in $\mathfrak{F}$ and are assigned the same probabilities by P as they are by P_i. Recalling Proposition 17, we can rewrite this using Kolmogorov notation as follows[17]:

$$\mathfrak{F}_i \subseteq \mathfrak{F}^{(\pi_i)} \text{ and } P^{(\pi_i)} = P_i \text{ on } \mathfrak{F}_i$$

We hence have 'agreement' whenever the component spaces of $(\Omega_1 \times \Omega_2, \mathfrak{F}, P)$ *contain*, respectively, the givens $(\Omega_1, \mathfrak{F}_1, P_1)$, $(\Omega_2, \mathfrak{F}_2, P_2)$. It turns out that containment is the best we can do in general. Nevertheless, *identity* is in fact possible in the case $\mathfrak{F}_1 = \mathfrak{F}_2 = \mathbb{R}$. A proof of this result, making use of the topological properties of $\mathbb{R}$, is offered in the Appendix.

Beyond Pairs

The construction above can be extended by induction to any finite number of spaces, without any problem. We are in such a case interested in any space $(\Omega_1 \times ... \times \Omega_n, \mathfrak{F}, P)$ that satisfies:

$$\mathfrak{F} =_{df} \sigma(\pi_1, ..., \pi_n)$$

$$\text{and} \quad P(\pi_i^{-1}(A_i)) = P_i(A_i), \quad (A_i \in \mathfrak{F}_i) \tag{2.18}$$

Among such spaces, one can prove the existence and uniqueness of the *product* space that additionally satisfies:

$$P(A_1 \times ... \times A_n) = P_1(A_1)P_2(A_2)...P_n(A_n), \quad (A_i \in \mathfrak{F}_i) \tag{2.19}$$

The analogous construction is possible for *countable* families of spaces[18]. First, to state the equivalent of (2.19) we need the following definition:

Definition 22. Let $(\Omega_i : i \in \mathbb{N})$ be a countable family of sets. A set $A \subseteq \prod_{i\in\mathbb{N}} \Omega_i$ is a *rectangle set* iff it is a product $\prod_{i\in\mathbb{N}} A_i$ of sets $A_i \subseteq \Omega_i$, where only a finite number of these containments are not identities. Formally:

$$A \text{ rectangle} \;\Leftrightarrow\; A = \prod_{i=1}^{n} A_i \times \prod_{j>n} \Omega_j, \qquad (\text{some } n)$$

$$\text{or equivalently } A \text{ rectangle} \;\Leftrightarrow\; A = \bigcap_{i\in I} \pi_i^{-1}(A_i), \qquad (\text{some finite } I)$$

We can now state the Theorem in question:

[17] We require that the pre-image ('encoding') of any set in $\mathfrak{F}_i$ be in $\mathfrak{F}$. This can be precisely stated as $A_i \in \mathfrak{F}_i \Rightarrow A_i \in \{A \mid \pi_i^{-1}(A_i) \in \mathfrak{F}\}$, or equivalently, $\mathfrak{F}_i \subseteq \mathfrak{F}^{(\pi_i)}$.

[18] It is also possible, under certain conditions, for uncountable families. We omit this discussion since we lack the technical notions required even to state the theorems in question.

Theorem 5 (Countable Product Space)**.** *Let* $((\Omega_i, \mathfrak{F}_i, P_i) : i \in \mathbb{N})$ *be a countable collection of probability spaces and let* $\Omega =_{df} \prod_i \Omega_i$*. Now define* π_i *to be the projections of* Ω *onto its component spaces and define the* product σ-algebra *over* Ω *as follows:*

$$\mathfrak{F} =_{df} \sigma(\pi_i : i \in \mathbb{N}) = \sigma(\{\pi_i^{-1}(A_i) \mid A_i \in \mathfrak{F}_i\})$$

There exists a unique choice of P *over* $(\Omega, \mathfrak{F})$ *that satisfies the following:*

$$P(\bigcap_{i \in I} \pi_i(A_i)) = \prod_{i \in I} P_i(A_i), \qquad (I \text{ finite}, A_i \in \mathfrak{F}_i) \tag{2.20}$$

Proof. Nontrivial, omitted, can be found in [15]. Uses techniques different to the analogous result for finite products. □

Remark. In the finite case, the product σ-algebra was understood variably either as the least σ-algebra containing all products $\prod_i A_i$, $A_i \in \mathfrak{F}_i$, or as the least σ-algebra that makes all projections measurable. In the countable case, the former definition becomes problematic, but the latter definition generalises painlessly.

Spaces with $\mathbb{R}^n$ as a set of Elementary Outcomes

Precisely as in the case of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the foundation for the construction of probability measures over $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is the notion of *n-dimensional Lebesgue measure*, a generalisation of the notion of 'volume'.

In this section, we will outline the key steps in the construction of n-fold Lebesgue measure, which is a straightforward abstraction of the 1-dimensional case. We will also use the results of the previous section to establish that the n-fold Lebesgue measure over $[0,1]^n$ agrees with the measure obtained by taking the product of n copies of $([0,1], \mathcal{B}([0,1]), Leb)$.

Definition 23. A **half-ray** in $\mathbb{R}^n$ is denoted by $L_{a_1 \ldots a_n}$ and is given by:

$$L_{a_1 \ldots a_n} =_{df} (-\infty, a_1) \times (-\infty, a_2) \times \ldots \times (-\infty, a_n), \quad (a_i \in \mathbb{R}) \tag{2.21}$$

Proposition 24.

$$\sigma(\{\textit{open sets in } \mathbb{R}^n\}) = \sigma(\{\textit{half-rays in } \mathbb{R}^n\})$$

Proof. We readily observe that half-rays are open sets. It now suffices to prove that open sets in $\mathbb{R}^n$ can be countably generated from the set of half-rays. By standard results, any open subset is a countable union of open 'hypercubes' of the form:

$$\prod_{1 \leq i \leq n} (a_i, b_i), \quad 0 < a_i, b_i \leq 1$$

We observe that:

$$\prod_i [a_i, b_i) = L_{b_1 \ldots b_n} - L_{a_1 b_2 \ldots b_n} - L_{b_1 a_2 \ldots b_n} - \ldots - L_{a_1 a_2 \ldots b_n}$$

and also that the singleton $\{(a_1, ..., a_n)\}$ is given by:

$$\{(a_1, ..., a_n)\} = \left(\bigcap_m L_{a_1 + \frac{1}{n}, ..., a_n + \frac{1}{m}} \right) \setminus L_{a_1 ... a_n}$$

which completes the proof since:

$$\prod_i (a_i, b_i) = \left(\prod_i [a_i, b_i) \right) \setminus \{(a_1, ..., a_n)\} \qquad \square$$

Theorem 6 (Existence of n-dimensional Lebesgue Measure)**.** *There exists a unique probability measure Leb_n on $([0,1]^n, \mathcal{B}([0,1]^n))$ such that:*

$$Leb_n(L_{a_1 ... a_n}) = \prod_{i=1}^{n} a_i, \qquad (0 \leq a_i \leq 1) \tag{2.22}$$

In later chapters where there is no risk of comfusion, we denote Leb_n by Leb.

Proof. The proof of existence is analogous to that of the one-dimensional result and can be found in [15]. The proof of uniqueness follows from Uniqueness Lemma since the set of half-rays in $\mathbb{R}^n$ is a π-system that generates $\mathcal{B}([0,1]^n)$. $\square$

This establishes the existence of the space $([0,1]^n, \mathcal{B}([0,1]^n), Leb_n)$, as required. The following lemma shows that this space is identical to the product space $([0,1]^n, \mathcal{B}([0,1]) \times ... \times \mathcal{B}([0,1]), Leb \times ... \times Leb)$:

Theorem 7. $\mathcal{B}(\mathbb{R}) \times ... \times \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^n)$

Proof. That $\mathcal{B}(\mathbb{R}^n) \subseteq \mathcal{B}(\mathbb{R}) \times ... \times \mathcal{B}(\mathbb{R})$ is trivial since the former is generated by n-dimensional half-rays (by Proposition 24) whereas the latter is generated by products of 1-dimensional Borel sets (by definition) and clearly n-dimensional half-rays are special cases of such products:

$$L_{a_1 ... a_n} =_{df} (-\infty, a_1) \times ... \times (\infty, a_n)$$

In the other direction, we need to use the following lemma:

Lemma. *The inverse image of a Borel set under a continuous map is Borel.*

Proof of Lemma. Consider two Borel spaces $(\Omega_1, \mathcal{B}(\Omega_1))$, $(\Omega_2, \mathcal{B}(\Omega_2))$, where

$$\mathcal{B}(\Omega_i) =_{df} \sigma(\mathcal{T}_1)$$

Assuming $u : \Omega_1 \to \Omega_2$ continuous, we are precisely required to show that

$$\mathcal{B}(\Omega_2) \text{ is contained in } \{A \subseteq \Omega_2 \mid u^{-1}[A] \text{ Borel in } \Omega_1\} \tag{2.23}$$

This containment follows by observing that $\{A \mid u^{-1}[A] \in \mathcal{B}(\Omega_1)\}$ is in fact a σ-algebra that contains $\mathcal{T}_2$: it contains $\mathcal{T}_2$ by the definition of continuity; it is a σ-algebra because the pre-image operator commutes with countable set operations and $\mathcal{B}(\Omega_1)$ is closed under countable set operations. $\square$

Finally, the Lemma applied to the continuous mapping π_i yields:

$$\text{for all } A_i \in \mathcal{B}(\mathbb{R}),\ \pi_i^{-1}(A_i) \in \mathcal{B}(\mathbb{R}^n)$$

$$\therefore\ \mathcal{B}(\mathbb{R}) \times ... \times \mathcal{B}(\mathbb{R}) =_{df} \sigma(\{\pi_i^{-1}(A_i) \mid A_i \in \mathcal{B}(\mathbb{R})\}) \subseteq \mathcal{B}(\mathbb{R}^n)$$

as required. □

Since the σ-algebras match, it follows that the measures $Leb \times ... \times Leb$ and Leb_n also match, since they agree on the π-system of half-rays by (2.19).

Remark. We will be using the lemma encountered above elsewhere in this dissertation as well. It is a very standard result in measure theory and is usually expressed by the statement 'continuous functions are Borel', where a *Borel* function is simply a measurable function between two Borel spaces:

Definition 24. Consider two Borel spaces $(\Omega_1, \sigma(\mathcal{T}_1))$, $(\Omega_2, \sigma(\mathcal{T}_2))$ and a probability function $u : \Omega_1 \to \Omega_2$. Then u is *Borel measurable* or *Borel* iff it is $\mathcal{T}_2/\mathcal{T}_1$-measurable.

2.3 Applying and Interpreting Probability Spaces

The first two examples are of special importance, since either of them can be thought of as a way of deriving the axioms of *finite* probability spaces from basic intuitions about counting and chance events[19].

Sequences of trials (frequentist interpretation)

Our first example of a setup that can be modelled by a *finite* probability space is known as the *finitistic frequentist interpretation* of the axioms and was the preferred interpretation of Kolmogorov at the time of writing of the Grundbegriffe. I summarily but faithfully reproduce his views here ([6, p.3]). It essentially concerns the modelling of repeatable experiments under controlled conditions (idealised laboratory experiments).

We study a definite set of events that could take place as a result of the establishment of an assumed complex of conditions, $\mathfrak{C}$, which allows of any number of repetitions. Then:

- $\Omega = \{\xi_1, \xi_2, ...\}$ is set equal to the *countable* set of all possible variants of the outcome of the experiment that we consider *à priori* possible.
- A set $A \subseteq \Omega$ is in $\mathfrak{F}$ iff it can be defined in any way that allows us to assert in a definite manner whether a given outcome belongs to A or not.

[19] Moreover, any space equipped with a finite σ-algebra where all the probabilities are rational can be brought in isomorphy to either of them.

- Under certain conditions, which Kolmogorov abstains from investigating in the Grundbegriffe, we can set $P(A)$ equal to a real number such that if m is the number of occurrences of A after a very large[20] number n of trials, then $\frac{m}{n}$ is very close to $P(A)$.

Having defined in this intuitive manner a space $(\Omega, \mathfrak{F}, P)$, we are in a position to derive that Axioms I-V hold of this space. Clearly Ω is well-defined, so is in $\mathfrak{F}$. So is any set defined from previously well-defined sets via countable[21] unions, intersections and differences. Hence Axioms I and II hold. Moreover, clearly $0 \leq m/n \leq 1$ and, since Ω occurs always, $P(\Omega) = n/n = 1$, which yields Axioms III and IV. Finally, if A and B are mutually exclusive and m_1, m_2 are the numbers of their respective occurrences in a sample of n trials, then clearly their disjunction has occurred precisely $m_1 + m_2$ times, yielding Axiom V.

This leaves Axiom VI, which clearly holds trivially if Ω is finite. Nevertheless, Kolmogorov's notation "$\Omega =_{df} \{\xi_1, \xi_2, ...\}$" also leaves open the possibility that Ω be countably infinite, in which case some separate argument is required to motivate Axiom VI. It is debatable whether or not it is possible to provide such an argument in this purely finitistic context of relative frequencies. Kolmogorov in 1933 thinks it "almost impossible" ([6, p.15]):

> Since the new axiom is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning, as has been done, for example, in the case of Axioms I -V [...].

Remark. The relative frequency interpretation also serves to motivate the crucial definition of elementary conditional probability as a reflection of the fact that in conditioning upon an event A, we are essentially focusing our interest in the *subsequence* of trials where A was indeed observed.

Single choice out of an urn (logical interpretation)

According to the rival *logical interpretation* of probability theory, a *finite* probability space can be defined only if one correctly identifies a fine-grained enough description of the problem so that elementary outcomes are in all respects symmetric, hence *equiprobable*. Such a description is meant to be reached via a *logical* breakdown of the outcomes of the experiment at hand into their finest constituents.

Once such a fine-grained description has been achieved, it can brought in isomorphy with an urn with n balls, for some natural number n. In this case

[20] The problem of how large 'very large' ought to be did not escape Kolmogorov nor his contemporaries. By some it was in fact viewed as a grave disadvantage of the frequentist foundations for probability. Taking a more modern standpoint, one would perhaps be inclined to view this problem not as a foundational problem of mathematics, but rather of science in general, very much akin the problem of accurate measurement in physics. As such, it cannot be claimed to pose a greater peril for probability theory than it does for any other scientific discipline.

[21] Kolmogorov only considers finite closure at this point in the text (see *Remark on Kolmogorov's Approach*, earlier).

$\Omega = \{\xi_1, ..., \xi_n\}$, $\mathfrak{F} = \mathcal{P}(\Omega)$. We posit that a random choice can be made out of that urn. By symmetry then we consider the selection of a certain ball to be equiprobable with the selection of any other. This precisely implies that $P(\xi_i) = 1/n$.

We then let some balls have an attribute that the others do not[22]. Intuitively, the probability of an attribute A coming up ought to be equal to

$$P(A) =_{df} \frac{\sharp A}{\sharp \Omega} = \frac{\sharp A}{n} \tag{2.24}$$

where $\sharp A$ counts the number of elements in the set A. This agrees with the calculations made via Axiom V and the values of P on the singletons:

$$\text{let } m =_{df} \sharp A, \text{ then} \quad P(A) = P(\{\xi_{i_1}, \xi_{i_2}, ..., \xi_{i_m}\}) = \sum_{j=1}^{m} P(\{\xi_{i_j}\}) = \frac{m}{n}$$

In this context, Axioms I and II hold by construction, whereas Axioms III, IV and V can be derived by the following simple counting arguments:

$$P(\Omega) = \frac{\sharp \Omega}{\sharp \Omega} = 1, \quad 0 \leq P(A) = \frac{\sharp A}{\sharp \Omega} \leq 1 \quad \text{and}$$

$$P(A \cup B) = \frac{\sharp (A \cup B)}{n} = \frac{\sharp A + \sharp B - \sharp (A \cap B)}{n} = P(A) + P(B) - P(A \cap B)$$

Remark. The definition of elementary conditional probability is sensible in this context, too. For readability, we denote the extensions of each attribute by the respective first letter. Now, what is the probability that a randomly chosen ball from a certain urn be wooden, given that it is black? The formulation of the question presupposes that in this particular experiment we have somehow been assured that the outcome *will* be a black ball. The space we are sampling out of is therefore not really Ω, but rather B, whereas the attribute 'wooden' no longer refers to W but to its restriction to black balls. These two observations together with (2.24) motivate the definition of conditional probability:

$$\text{assuming } B \neq \emptyset, \quad P_B(W) = \frac{\sharp (B \cap W)}{\sharp B} = \frac{\sharp (B \cap W)/n}{\sharp B/n} = \frac{P(B \cap W)}{P(B)}$$

Logical Interpretation for Infinite Spaces - a discussion[23]

Both the frequentist and logical interpretations offered above fail to apply to spaces where Ω is *continuous*[24]. Nevertheless, the logical interpretation has a

[22] Care must be taken in practice that this does not affect the random selection (for instance, if randomness is ensured by blindfolding the person who picks, attributes of colour can be brought into the discussion but attributes of shape cannot).

[23] This discussion largely repeats certain arguments made in the Introduction.

[24] In fact, even the countable case is problematic - the frequentist interpretation fails to incorporate Axiom VI, whereas the logical interpretation directly contradicts it, since there exists no space with a countable set of equiprobable elementary events: in any such space, by countable additivity $P(\Omega)$ should equal either ∞ or 0.

particularly simple analogue for the continuous case, which is of relevance to Bertrand paradoxes. We review this analogue here.

One assumes that a real number $x \in [a, b]$ is said to be chosen *completely* or *uniformly at random* if the probability that it ends up in any interval is equal to the length of that interval[25]. This of course corresponds to the *Lebesgue* measure on $[0, 1]$, otherwise known in this context as the *uniform* measure.

Therefore the task of the logical probabilist now becomes one of reparameterising the problem, until the parameter 'sampled' is sampled 'uniformly at random'. It is however questionable whether this task can be accomplished by the same 'logical' reasoning that is usually employed in the finite case. In particular, the following basic intuition about symmetric outcomes now fails:

> if we have two descriptions of the set of elementary outcomes that are in bijective correspondence between them, then if in one of them elementary outcomes are equiprobable, this also holds of the other.

That this fails in the continuous case can be witnessed via even very simple bijections on $[0, 1]$, such as the map $f(x) = x^2$. This failing lies at the very core of Bertrand paradoxes, as we will explain in Chapter 3.

Two choices out of an urn with replacement

Let us now return to the task of modelling simple experiments and try to represent the choice of two balls, *with replacement*, from an urn: the experimenter picks a ball at random, registers its attributes, then replaces it and repeats once more. Assume $(\Omega, \mathfrak{F}, P)$ models a single choice out of an urn $\Omega = \{\xi_1, ..., \xi_n\}$ as before. For two choices we then must have:

$$\Omega' =_{df} \Omega \times \Omega, \quad \mathfrak{F}' = \mathcal{P}(\Omega)$$

Moreover, the following must hold true of P', since by symmetry any pair (ξ_i, ξ_j) is equally likely:

$$P'(A \times B) =_{df} \frac{\sharp(A \times B)}{\sharp\Omega} = \frac{(\sharp A)(\sharp B)}{n^2} = P(A)P(B)$$

This identifies $(\Omega', \mathfrak{F}', P')$ with the product space[26] of two copies of $(\Omega, \mathfrak{F}, P)$. Note that we could have equally well derived this by insisting that the second

[25] There are several ways to motivate this choice. One is to insist that the probability be invariant with respect to translations, expressing a certain symmetry with respect to the position of x. This yields the Lebesgue measure. Another is to view decimal expansions as sequences of random selections from $\{0, 1, ..., 9\}$, where each digit has equal probability. As we will establish at the end of this section, this representation also yields the Lebesgue measure.

[26] This example gives us an opportunity to make an important remark: Kolmogorov's choice of formalisation picks out the notion of an *experiment* as the most basic notion of informal probability talk and uses it as the cornerstone of the formal theory as well. This is also remarked by Prof. Williams in the opening sentence of the second chapter of his textbook as follows ([15, p.23]):

> A model for an experiment involving randomness takes the form of a *probability triple* $(\Omega, \mathfrak{F}, P)$ [...].

choice be independent of the first. We will demonstrate this alternative route in dealing with the

Two choices out of an urn without replacement

Let us now model the experiment where a choice of two balls **without replacement** from the urn is being made at random: the experimenter randomly chooses a ball, registers its attributes and then repeats without replacing the ball. The first choice determines what urn the second choice is made out of.

Naturally, we have to ban pairs of the form (ξ_i, ξ_i). We can either do that explicitly by leaving them out of our set of elementary outcomes, or do it implicitly by giving them zero probability. We take the latter route and work over the full Cartesian product $\Omega \times \Omega$ and its power set. The choice of P'' can now be dictated via the use of conditional probability from elementary intuitions, namely that symmetry holds for each choice:

$$\text{1st choice is symmetric:} \quad P''(\text{1st choice is } \xi) = \frac{1}{n}$$

$$\text{2nd choice is symmetric:} \quad P''_{\text{1st choice is } \xi}(\text{2nd choice is } \xi') = \begin{cases} \frac{1}{n-1}, \text{ if } \xi \neq \xi' \\ 0, \text{ if } \xi = \xi' \end{cases}$$

The uniqueness of conditional probability (Proposition 4) ensures us that there exists only one probability assignment on the singletons that satisfies both these constraints:

$$\begin{aligned} P''(\xi, \xi') &= P''(\text{1st choice is } \xi) P''_{\text{1st choice is } \xi}(\text{2nd choice is } \xi') \\ &= \begin{cases} \frac{1}{n(n-1)}, \text{ if } \xi \neq \xi' \\ 0, \text{ if } \xi = \xi' \end{cases} \end{aligned}$$

Since this is a finite space, this assignment on the singletons suffices to uniquely induce the full probability measure. Now, typically[27] in this situation we are

This principle can be seen at work in the fact that the probability space $(\Omega, \mathfrak{F}, P)$ is related, but not the same with $(\Omega', \mathfrak{F}', P')$. This reflects the fact that a single choice out of an urn is different as an *experiment* than two choices out of an urn, although the experimental setups are the same (namely, the urn).

This is a point of importance, since there is no shortage of alternative formalisations of probability theory, in which different basic notions are employed: for instance, that of a random sequence, a random source or a degree of belief.

[27] In teaching of probability theory in secondary education it is customary to ask this question for particular choices of A_1 and A_2. The student is expected to concoct a separate argument for each such choice. For instance, assume that we have an urn with n balls, n_1 of which are black, n_2 red and n_3 green (so $n_1 + n_2 + n_3 = n$). Assume we now pick a ball without replacement. What is the probability that the second ball is red given that the first one was black? We expect this answer to be $\frac{n_2}{n-1}$. Indeed, letting A_1 be the set of black balls and A_2 the set of red ones, $A_1 \cap A_2 = \emptyset$, so (2.25) becomes:

$$\frac{n(A_1)n(A_2) - n(A_1 \cap A_2)}{n(A_1)(n-1)} = \frac{n_2}{n-1}$$

asked to calculate the probability that the second choice has a certain attribute A_2 given that the first had attribute A_1. Attributes A_1 and A_2 are represented formally via their extensions, as follows:

$$\text{1st choice is in } A_1 =_{df} \{(\xi_i, \xi_j) \mid \xi_i \in A_1\} =_{df} X$$

$$\text{2nd choice is in } A_2 =_{df} \{(\xi_i, \xi_j) \mid \xi_j \in A_2\} =_{df} Y$$

Therefore, by an easy calculation which we omit, we have (provided $A_1 \neq \emptyset$):

$$P''_{\textit{1st choice is in } A_1}(\textit{2nd choice is in } A_2) = \frac{n(A_1)n(A_2) - n(A_1 \cap A_2)}{n(A_1)(n-1)} \tag{2.25}$$

Denumerable coin tossing

We now move to a different kind of experiment:

> a biased coin with probability of heads equal to p, where $0 < p < 1$,
> is tossed repeatedly until tails turns up.

We 'reset' our notation and denote by Ω the underlying set of possible outcomes, which in this case is given by

$$\Omega = \{T, HT, HHT, HHHT, ...\} \cup \{HHHH...\}.$$

In particular it must contain the infinite sequence 'HHHH...', since in principle (and in literature[28]) the coin might for ever refuse to turn up tails. We give this sequence the special name H^∞.

A fully formal approach would express Ω in terms of countable products of single tosses. For a while, we allow ourselves a significant head start and rely on our intuition[29] to explicitly assign probabilities on singletons of Ω:

$$P(\{x\}) =_{df} \begin{cases} 0, \text{ if } x = H^\infty \\ (1-p)\,p^{n-1}, \text{ otherwise} \end{cases} \qquad (x \in \Omega) \tag{2.26}$$

where n is the length of the sequence x, a finite number whenever $x \neq H^\infty$. It follows that

$$\begin{aligned} P(\Omega) &= P(\{H^\infty\}) + \sum_{x \in \Omega \setminus \{H^\infty\}} P(\{x\}), \text{ by countable additivity} \\ &= 0 + \sum_{n=1}^{\infty} (1-p)p^{n-1}, \text{ by definition of } P \\ &= 1, \text{ by the summation of the geometric series.} \end{aligned}$$

[28] In the well-known play *Rosencrantz and Guildenstern are Dead* by Tom Stoppard, Guildenstern, on his way to assassinate Hamlet, loses a substantial amount of money on account of a certain coin for ever refusing to turn up heads. The character interprets this alarmingly bad luck as a sign that he is soon meant to die. Indeed, he is right. Just as such a sequence of die tosses can only be predetermined, so is the fate of Guildenstern, having been decided centuries ago when Shakespeare described his death in *Hamlet*.

[29] This is as formal as one would get in an applied probability course.

We hence have an assignment of probabilities on the singletons such that $P(\Omega) = 1$. From the results on spaces over countable sets of elementary outcomes, this uniquely defines a probability measure on $\mathcal{P}(\Omega)$.

Completing the gaps, we now demonstrate a more formal alternative, wherein the assignment (2.26) is derived from the independence of the tosses. To do so, we have to look at a much bigger space, which we now describe. First, let $(\Omega_0, \mathfrak{F}_0, P_0)$ denote a single toss of the biased coin:

$$\Omega_0 = \{H, T\}, \quad \mathfrak{F}_0 = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$$

$$P_0(\{H\}) = p, \quad P_0(\{T\}) = 1 - p, \qquad \text{for some } 0 < p < 1$$

We can now represent infinite sequences of tosses as elements of the cartesian product of countably many identical copies of Ω_0:

$$\Omega' =_{df} \{H, T\}^{\mathbb{N}} \cong (\mathbb{N} \to \{H, T\})$$

Naturally, we equip it with the respective product σ-algebra:

$$\mathfrak{F}' =_{df} \sigma(\{R \mid R \text{ is a rectangle set}\})$$

where *rectangle sets*, as explained in the section on product spaces, are finite intersections of inverse images of projections. Since the inverse image under the i'th projection map can only be one of the two sets S_i and S_i^c defined below:

$$S_i =_{df} \pi_i^{-1}(\{H\}) = \{s \in \Omega' \mid s(i) = H\}$$

$$S_i^c =_{df} \pi_i^{-1}(\{T\}) = \{s \in \Omega' \mid s(i) = T\}$$

then an arbitrary rectangle R is given by

$$R =_{df} \prod_{i \in Q} S_i \cap \prod_{j \in R} S_j^c, \quad (Q, R \subseteq \mathbb{N}, \text{ disjoint and finite})$$

We now factor in the independence of the tosses and the understanding that the probability of the i'th toss is constant, equal to p. Together these enforce the following requirement on our probability measure P':

$$P'\left(\prod_{i \in Q} S_i \cap \prod_{j \in R} S_j^c\right) = \prod_{i \in Q} P'(S_i) \prod_{j \in R} P'(S_j^c) = p^q (1 - p)^r \tag{2.27}$$

where $q = \sharp Q$ and $r = \sharp R$. Clearly, this requirement[30] precisely defines $(\Omega', \mathfrak{F}', P')$ to be the countable product space of countably infinite many copies of $(\Omega_0, \mathfrak{F}_0, P_0)$.

[30]The crucial application of Theorem 5 at this point means that there is no need in (2.27) to posit independence for infinitely many trials - just for any finite combination of trials. This is the difficulty that Borel had not addressed in his strong law of large numbers and which Kolmogorov eventually resolved using the extension theorems.

We presently compare $(\Omega, \mathfrak{F}, P)$ with $(\Omega', \mathfrak{F}', P')$. First, we need to represent the singletons of Ω in Ω'. This is trivial in the case of H^∞:

$$\{H^\infty\} \text{ is given by } \bigcap_{i\in\mathbb{N}} S_i$$

Moreover, it is clear that any finite sequence is represented in $\mathfrak{F}'$ by the set of all its possible continuations. In this spirit, the singletons of $\Omega \setminus \{H^\infty\}$ are represented in $\mathfrak{F}'$ as follows:

$$\{HHH...T\} \text{ is given by } \bigcap_{i=1}^{n-1} S_i \cap S_n^c, \quad \text{where } n \text{ is the length of } HH...T.$$

We now check whether the probabilities of these sets in $(\Omega', \mathfrak{F}', P')$ match our intuitive assignment (2.26) on the singletons of Ω. First,

$$P'(\{HHH...T\}) = P(S_n^c) \prod_{i=1}^{n-1} P(S_i) = (1-p)p^n, \text{ directly via (2.27)}$$

yielding one branch of (2.26). The other branch is also easy to recover:

$$\begin{aligned} P'(H^\infty) &= P'(\lim_{n\to\infty} \bigcap_{i=1}^{n} S_i), \text{ since } \{H^\infty\} \text{ is given by } \bigcap_{i\in\mathbb{N}} S_i \\ &= \lim_{n\to\infty} P'(\bigcap_{i=1}^{n} S_i), \text{ by monotone convergence (Proposition 3)} \\ &= \lim_{n\to\infty} p^n, \text{ by (2.27)} \\ &= 0, \text{ since } 0 < p < 1. \end{aligned}$$

This completes the derivation of (2.26) from the independence of tosses, as promised. We have gained much more, though: whereas $(\Omega, \mathfrak{F}, P)$ was tailored to answer one specific question, the space $(\Omega', \mathfrak{F}', P')$ can answer *any* sensible question we wish to ask about an infinite number of coin tosses. For instance:

Theorem 8 (Borel)**.** *Almost surely H will occur infinitely often.*

Remark. Naturally we insist on the assumption that $0 < p < 1$.

Proof. Consider the singleton of any sequence s of tosses with a finite number of occurrences of H. Then from a certain toss onwards, say the m'th toss, only T ever comes up. This precisely means that:

$$\{s\} \subseteq \bigcap_{i=m}^{\infty} S_i^c, \quad \text{therefore} \quad P(\{s\}) \leq P\left(\bigcap_{i=m}^{\infty} S_i^c\right) = 0$$

again by Proposition 3. Moreover, the set of all such s is countable, since

$$\{s \mid H \text{ occurs finitely many times}\} = \bigcup_{n\in\mathbb{N}} \{s \mid H \text{ does not occur after } n\text{'th toss}\}$$

and the sets under the union in the RHS are finite. So by countable additivity

$$P'(H \text{ occurs finitely many times}) = 0,$$
$$\therefore \qquad P'(H \text{ occurs infinitely often}) = 1. \qquad \square$$

Denumerable coin tossing and Lebesgue Measure

If, throughout the above example, we replace the biased coin with an unbiased dekahedron (ten-sided die), we can then think of each infinite sequence of tosses to represent the decimal expansion of a certain real number $x \in [0, 1]$. We can make this mapping bijective if we ban sequences that end in an infinite repetition of 9's (they form a countable set with zero probability eitherway).

It is then easy to show that the end-product of our construction of $(\Omega', \mathfrak{F}', P')$ above is equivalent to the *Lebesgue measure* over $[0, 1]$. For instance:

$$[0.13, 0.14] \text{ is represented in } \mathfrak{F}' \text{ by } \pi_1^{-1}(1) \cap \pi_2^{-1}(3)$$

$$\therefore \quad P'([0.13, 0.14]) = P'(\pi_1^{-1}(1))P'(\cap\pi_2^{-1}(3)) = \frac{1}{100} = Leb([0.13, 0.14])$$

This argument in fact works for any interval whose endpoints admit of a finite decimal expansion. This together with the following simple proposition, suffices to prove that $P' = Leb$ for any Borel set, via the Uniqueness Lemma:

Proposition 25. *Let $\mathcal{I}$ be the set of intervals whose endpoints admit of a finite decimal expansion. Then $\mathcal{I}$ is a π-system that generates the Borel sets in $[0, 1]$:*

$$\sigma(\mathcal{I}) = \mathcal{B}[0, 1]$$

Proof. Clearly the reals that admit of a finite decimal expansion form a dense set in $[0, 1]$, as this is equivalent to saying that the decimal expansion representation of the reals exists. So for any interval $[a, b]$ we can found a countable sequence of intervals that tends to it, in the sense that $[a, b]$ is the infimum of the sequence. This proves that $\sigma(\mathcal{I})$ contains all intervals and hence also all Borel sets. $\square$

Remark. It was precisely in the context of this argument that Emile Borel was first motivated to introduce the property of countable additivity.

What about spaces over $\Omega = \mathbb{R}^n$ other than the Lebesgue measure?

As we mentioned earlier, spaces over $\mathbb{R}^n$ are usually defined via the concept of a distribution function, which is the topic of the next subsection. For concrete examples of such spaces, the reader is invited to wait until Chapter 3, where Bertrand paradoxes are investigated.

2.4 Random Variables and Spaces over $\mathcal{B}\mathbb{R}^n$

We will now turn our attention to probability functions *into the reals*. We have explained that a probability function $u : \Omega \to \Omega'$ is useful in that it allows us to reason about $(\Omega, \mathfrak{F}, P)$ *indirectly*, by investigating the induced space $(\Omega', \mathfrak{F}^{(u)}, P^{(u)})$. In particular, probability functions into the reals allow us to take advantage of topological methods, provided the topology of the reals can is contained in the induced field $\mathfrak{F}^{(u)}$. Equivalently, the Borel σ-algebra over the reals must be contained in $\mathfrak{F}^{(u)}$:

$$\mathcal{B}(\mathbb{R}) \subseteq \mathfrak{F}^{(u)}$$

This motivates[31] the following definition:

Definition 25. A probability function $u : \Omega \to \mathbb{R}$ carried by the probability space $(\Omega, \mathfrak{F}, P)$ is a *random variable* iff $\mathcal{B}(\mathbb{R}) \subseteq \mathfrak{F}^{(u)}$.

Proposition 26. *The following are equivalent:*

A. The probability function u is a random variable.

B. The probability function u is $\mathcal{B}(\mathbb{R})/\mathfrak{F}$-measurable.

C. For every $a \in \mathbb{R}$, $(-\infty, a) \in \mathfrak{F}^{(u)}$.

D. For every $a \in \mathbb{R}$, $(-\infty, a] \in \mathfrak{F}^{(u)}$.

Proof. Claim (A) is equivalent to (B) by definition (see *Remark on Measurability* in Section 2.1). The equivalence of (C) with (A) follows from Proposition 22 and the equivalence of (D) with (A) from the remark that follows it. □

Remark. On the basis of the previous proposition we will restrict our attention to Borel sets only in this section,

$$\text{working with } (\mathbb{R}, \mathcal{B}(\mathbb{R}), P^{(u)}), \quad \text{as opposed to } (\mathbb{R}, \mathfrak{F}^{(u)}, P^{(u)})$$

We take this step since, in dealing with distribution functions, our purpose is precisely to investigate probability measures over the Borel σ-algebra on the reals. In modern probability theory, this restriction is presupposed straight from the onset, by defining a random variable to be a $\mathcal{B}(\mathbb{R})/\mathfrak{F}$-measurable function.

[31] A less formal motivation arises by thinking of a random variable as roughly capturing the concept of a 'measuring device', or an 'observable'. Consider an arbitrarily complex underlying process that yields *measurements* via some device. It would then be desirable that the set of possible measurements be (a subset of) a complete ordered field, namely $\mathbb{R}$, and that we can always ask the question "what is the probability that the measured quantity is below a certain value a?". These are precisely the two formal requirements we demand of a probability function so that it be called a 'random variable', as described in (C) of Proposition 26.

Half-rays[32] have two useful properties: firstly, they generate the Borel sets of $\mathbb{R}$ (Proposition 22). Secondly, their probabilities can be naturally given in a functional form, via a *distribution function*:

Definition 26. The *distribution function* $F^{(x)}$ of a random variable x is:

$$F^{(x)}(a) =_{df} P^{(x)}((-\infty, a)), \quad (a \in \mathbb{R}) \tag{2.28}$$

Remark. Distribution functions are often defined over the extended real field which includes the points $\infty, -\infty$, under the understanding that $(-\infty, -\infty) =_{df} \emptyset$ and $(-\infty, \infty) =_{df} \mathbb{R}$. It follows directly from the properties of measure that for any random variable x, $F^{(x)}(\infty) = 1$ and $F^{(x)}(-\infty) = 0$.

Proposition 27. *Let x be an r.v. and F its distribution function. Then*[33]*:*

$$F : \mathbb{R} \to [0, 1] \tag{2.29}$$

$$F \text{ is non-decreasing} \tag{2.30}$$

$$lim_{x \to -\infty} F(x) = 0 \text{ and } lim_{x \to \infty} F(x) = 1 \tag{2.31}$$

$$F \text{ is continuous on the left.} \tag{2.32}$$

Proof. By Proposition 26 all half-rays have probabilities, which proves (2.29). Moreover, for all $a, b \in \mathbb{R}$:

$$a < b \;\Rightarrow\; (-\infty, a) \subseteq (-\infty, b) \;\Rightarrow\; P(-\infty, a) \subseteq P(-\infty, b) \;\Rightarrow\; F(a) \leq F(b)$$

which proves (2.30).

Now for any sequence of reals $a_1, a_2, ...$ diverging to ∞, we have that

$$(-\infty, a_1) \subseteq (-\infty, a_2) \subseteq ... \quad \text{and} \bigcup_{n \in \mathbb{N}} (a_n, \infty) = \mathbb{R}$$

so by monotone convergence

$$\lim_{n \to \infty} F(a_n) = \lim_{n \to \infty} P((-\infty, a_n)) = P(\mathbb{R}) = 1$$

Entirely analogously monotone convergence also yields that

$$\lim_{n \to \infty} a_n = -\infty \;\Rightarrow\; \lim_{n \to \infty} F(a_n) = 0$$

$$\text{and} \qquad \lim_{n \to \infty} a_n = b \;\Rightarrow\; \lim_{n \to \infty} (F(b) - F(a_n)) = 0$$

which proves (2.31) and (2.32). □

[32] In some modern textbooks, closed half-rays $(-\infty, a]$ are used instead of open half-rays throughout the discussion in this section. Our choice of open half-rays seems to me neater since it emphasizes the fact that the same system of open sets generates both the topology and the Borel σ-algebra of $\mathbb{R}$. As we have explained both approaches are equivalent, although the choice of approach affects several technical details in certain proofs. We will take care to make note of such discrepancies.

[33] Had we been using closed half-rays instead, this proposition would still go through, except with right-continuity rather than left-continuity in (2.32).

In fact, not only are (2.29)-(2.32) *necessary* for a certain function to be the distribution function of a random variable, but they are also *sufficient*, as the following result, found in [15, p.34] establishes:

Theorem 9 (Skorokhod Representation)**.** *Let F be any function that satisfies (2.29)-(2.32). Then the function $X_F : [0,1] \to \mathbb{R}$ defined by:*

$$X_F(\omega) =_{df} \inf\{z : F(z) > \omega\}$$

is a random variable with distribution function F when understood as a probability function from the space $([0,1], \mathcal{B}([0,1]), Leb)$. We call X_F the Skorokhod Representation[34] of F.

Proof. In the Appendix. □

Therefore, any function that satisfies (2.29)-(2.32) is the distribution of *some* random variable, which validates the following choice of terminology:

Definition 27. A function F that satisfies (2.29)-(2.32) is called a *distribution.*

Corollary. *The following equation defines a bijective correspondence between probability measures P' on $(\mathbb{R}, \mathcal{B}\mathbb{R})$ and distribution functions F:*

$$F(a) =_{df} P'((-\infty, a)) \tag{2.33}$$

Proof. Let P' be an arbitrary probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then the following assignment:

$$F(a) =_{df} P'((-\infty, a))$$

uniquely defines a distribution F, namely the distribution function of the identity, viewed as a random variable.

Conversely, let F be an arbitrary distribution. Then consider $P^{(X_F)}$, the probability measure on $\mathbb{R}$ induced by the Skorokhod Representation[35] of F. By construction, the distribution of X_F is F, so $P^{(X_F)}$ satisfies (2.33):

$$P^{(X_F)}((-\infty, a)) = F(a)$$

On the other hand, there cannot be any *other* measure over $\mathcal{B}(\mathbb{R})$ that does, by Uniqueness Lemma. In particular, if F had initially been defined via (2.33) from an arbitrary probability measure P', then it must be that $P^{(X_F)} = P'$. □

Remark. This correspondence we have just established allows applied probability textbooks to encourage students to think of probability spaces directly in terms of distribution functions, rather than measure spaces, hence allowing the tools of calculus to be used in more familiar ground.

[34] Had we assumed that distributions are right-continuous, X_F would have to defined as $\inf\{z : F(z) \geq \omega\}$ for an analogous proof to go through. In any case, the two versions of X_F can be easily seen to agree almost everywhere.

[35] This measure is otherwise known as the *Stieltjes measure with respect to F* - see the discussion of the Skorokhod Representation Theorem that follows.

The Skorokhod Representation Theorem: a discussion

Theorem 9 ensures us that any random variable x carried by an *arbitrary* space $(\Omega, \mathfrak{F}, P)$, insofar as the investigation of the induced probabilities of Borel sets are concerned, can be equally well represented by the Skorokhod representation of the distribution function of x, which is a random variable carried by $([0,1], \mathcal{B}([0,1]), Leb)$. This explains the following important observation made by David Williams in [15, p.35]:

> It is in fact true that every experiment you will meet in this (or any other) course can be modelled via the triple $([0,1], \mathcal{B}[0,1], Leb)$.

Williams then hastens to add:

> However, this observation normally has only curiosity value.

In my opinion, it truly could be argued that Williams' initial observation "normally has only curiosity value" in a *philosophical* sense. Similarly its importance is little in a *practical*, *applied* sense; the correspondence between measures and distribution functions being instead the key tool in terms of applicability. However, the modelling 'omnipotence' of $([0,1], \mathcal{B}[0,1], Leb)$ is of grand *foundational* significance, especially in the context of geometric probability. Indeed, with some care one will observe that the existence and properties of all examples of continuous probability spaces contained in this dissertation flow from the existence and properties of Lebesgue measure.

On a technical note, it ought to be remarked that the bijective correspondence between probability measures over $\mathcal{B}(\mathbb{R})$ and distribution functions can be proved without the Skorokhod Representation Theorem, by properly amending the argument used for the proof of the existence of Lebesgue measure. Our approach, although less direct, is better suited to our discussion, since it presupposes the existence of Lebesgue measure and only does the necessary 'extra work', hence emphasizing the foundational importance of the existence of Lebesgue measure and avoiding the repetition of technical proofs.

For completeness, we make a note here of the notation and terminology employed in the absence of the Skorokhod Representation Theorem:

Definition 28 (Stieltjes Measure)**.** Let F be a distribution. Then consider the unique probability measure m_F over $\mathcal{B}(\mathbb{R})$ that satisfies the following:

$$m_F(\{z \in \mathbb{R} \mid z < a\}) =_{df} F(a) \tag{2.34}$$

This choice of measure is called the *Stieltjes measure with respect to* F.

In the terminology of Theorem 9, the Stieltjes measure with respect to F is the measure induced over $\mathcal{B}(\mathbb{R})$ by the Skorokhod Representation of F, ie:

$$m_F =_{df} P^{(X_F)}$$

Multidimensional Distribution Functions

Consider n random variables, x_1, ..., x_n carried by a probability space $(\Omega, \mathfrak{F}, P)$. We now view the tuple of the n random variables as a probability function from Ω to $\mathbb{R}^n$:

$$x : \omega \mapsto (x_1(\omega), x_2(\omega), ..., x_n(\omega)), \qquad (\omega \in \Omega,\ x_i(\omega) \in \mathbb{R})$$

We can now work in the induced space $(\mathbb{R}^n, \mathfrak{F}^{(x)}, P^{(x)})$. However we choose to restrict it to the Borel sets only, analogously to the one-dimensional case. We are allowed to do that since

Proposition 28. $\mathfrak{F}^{(x)} \supseteq \mathcal{B}(\mathbb{R}^n)$

Proof. It suffices to show that the n-fold half-rays are in $\mathfrak{F}^{(x)}$, since $\mathcal{B}(\mathbb{R}^n)$ is the least σ-algebra that contains them.

$$\begin{aligned} x^{-1}(L_{a_1 \ldots a_n}) &= \{\omega \mid x(\omega) =_{df} (x_1(\omega), ..., x_n(\omega)) \in L_{a_1 \ldots a_n}\} \\ &= \{\omega \mid x_i(\omega) \in (-\infty, a_i),\ 1 \leq i \leq n\} \\ &= \bigcap_{i=1}^{n} x_i^{-1}((-\infty, a_i)) \in \mathfrak{F} \end{aligned}$$ □

Remark. One is certainly tempted to use the methods employed in the section on product spaces to see in what way $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P^{(x)})$ is related to the spaces induced by each random variable on its own, again restricted to Borel sets, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P^{(x_i)})$. We remind the reader that the crucial conditions are two, one of *agreement* and one of *independence*, labelled (2.18) and (2.19) respectively:

Agreement. for A_i Borel, $P^{(x)}(\pi_i^{-1}(A_i)) = P^{(x_i)}(A_i), \quad (1 \leq i \leq n)$

Independence. for A_i Borel, $P^{(x)}(A_1 \times ... \times A_n) = \prod_{i=1}^{n} P^{(x_i)}(A_i)$

It is easy to see that agreement is guaranteed by the definition of x, since the composition of x with the projection π_i yields precisely x_i. On the other hand, clearly independence will not always hold and $P^{(x)}$ may very well not agree with the product measure $P^{(x_1)} \times ... \times P^{(x_n)}$, as we will see in Proposition 29 below. This is natural, since the existence of the underlying space $(\Omega, \mathfrak{F}, P)$ yields the expressive power to introduce correlations. It is also desirable since it provides a method of defining measures over $\mathcal{B}(\mathbb{R}^n)$ other than the product measure. Such a method was lacking in Section 2.2.

Proposition 29. *It is not always true that* $P^{(x)} = P^{(x_1)} \times ... \times P^{(x_n)}$.

Proof. Consider the probability space $([0,1], \mathcal{B}([0,1]), Leb)$ equipped with two random variables x_1, x_2. The requirement of independence demands that for any Borel sets A_1 and A_2,

$$\begin{aligned} & P^{(x)}(A_1 \times A_2) = P^{(x_1)}(A_1) P^{(x_2)}(A_2) \\ \Leftrightarrow \quad & P(x_1^{-1}(A_1) \cap x_2^{-1}(A_2)) = P(x_1^{-1}(A_1)) P(x_2^{-1}(A_2)) \end{aligned} \tag{2.35}$$

Now set $x_1(\omega) = x_2(\omega) = \omega$ and $A_1 = A_2 = [0, 0.5]$, in which case the LHS of (2.35) becomes:

$$\begin{aligned} P(x_1^{-1}(A_1) \cap x_2^{-1}(A_2)) &= P(x_1^{-1}(A_1)), \text{ since } x_1 = x_2 \text{ and } A_1 = A_2 \\ &= P(A_1), \text{ since } x_1 \text{ is the identity and } A_1 \subseteq [0,1] \\ &= 0.5 \end{aligned}$$

whereas by the same reasoning the RHS of (2.35) is equal to 0.25. □

Having briefly investigated the properties of the tuple function, we can now proceed to define the multidimensional analogue of distribution functions:

Definition 29. The following function is called the *n-dimensional distribution function of the random variables* x_1, ..., x_n:

$$F^{(x_1,\ldots,x_n)} =_{df} P^{(x)}(L_{a_1\ldots a_n}) \tag{2.36}$$

Proposition 30. *Let* $x_1, ..., x_n$ *be random variables and* F *their n-dimensional distribution function. Then:*

$$F : (\mathbb{R} \cup \{-\infty, \infty\})^n \to [0,1] \tag{2.37}$$

F *is non-decreasing in each variable* (2.38)

$$lim_{a_i \to -\infty} F(a_1, ..., a_n) = F(a_1, ..., a_{i-1}, -\infty, a_{i+1}, ..., a_n) = 0 \tag{2.39}$$

$$lim_{a_1,\ldots,a_n \to \infty} F(a_1, ..., a_n) = F(\infty, ..., \infty) = 1 \tag{2.40}$$

F *is continuous on the left in each variable* (2.41)

Proof. The proof is entirely analogous of that of Proposition 27. □

Definition 30. Any function $F : \mathbb{R}^n \to [0,1]$ that satisfies (2.37)-(2.41) is called an *(n-dimensional) distribution function.*

Theorem 10. *The following equation defines a bijective correspondence between probability measures* P' *on* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ *and n-dimensional distributions* F*:*

$$F(a_1, ..., a_n) =_{df} P'(L_{a_1\ldots a_n}) \tag{2.33}$$

We denote the measure P' *that corresponds to the distribution* F *by* m_F*.*

Proof. Again, the proof is analogous to the one-dimensional case. □

Probability Density Functions

Distribution functions become even more useful when they satisfy certain additional 'nicety' conditions, in which case calculus can be brought to bear on the computation of probability values. The result that follows presupposes an understanding of Lebesgue Integrability. However, it is best to place it here and invite the reader to revisit its proof after a proper treatment of the integral is provided in Section 2.5.

Proposition 31. *Let F be an n-dimensional distribution function. Assume that it is also the case that for some Borel-measurable function f:*

$$F(a_1...a_n) = \int_{L_{a_1...a_n}} f(t_1, ..., t_n) d_{Leb}(\mathbf{t}), \quad ((a_1, ..., a_n) \in \mathbb{R}^n) \tag{2.42}$$

Then for any Borel set A we have:

$$m_F(A) = \int_A f(\mathbf{t}) d_{Leb}(\mathbf{t})$$

where all integrals are understood in the Lebesgue sense.

Proof. We prove the result for the case $n = 1$ only. The multidimensional case can be proved analogously. Let μ be defined via:

$$\mu(A) =_{df} \int_A f(t) d_{Leb}(t)$$

Then μ is countably additive by Property II of Theorem 15. Moreover, since F is monotonically increasing, f is nowhere negative, therefore $\mu(A) \geq 0$. Finally, since $F(a) \to 1$ as $a \to \infty$, we get that:

$$\mu(\mathbb{R}) = \int_{-\infty}^{\infty} f(t) d_{Leb}(t) = 1$$

Therefore μ is a measure. But now observe that:

$$\begin{aligned} \mu((-\infty, a)) &=_{df} \int_{-\infty}^{a} f(t) d_{Leb}(t) \\ &= F(a), \text{ by hypothesis (2.42)} \end{aligned}$$

So μ is in fact identical to m_F on the π-system of half-rays and therefore is identical to it everywhere, by Uniqueness Lemma. □

Remark. In the 1-dimensional case, under suitable conditions, we may write

$$f =_{df} \frac{d}{da} F(a) \tag{2.43}$$

for the function f that appears under the Lebesgue integral in (2.42). In this case, the function is called the *probability density function* of F. The relationship between definition (2.42) and definition (2.43) is simple but not trivial and involves the Fundamental Theorem of Calculus for Lebesgue integrals. An equivalent discussion is possible for the n-dimensional case, using *partial derivatives*:

$$f =_{df} \frac{\partial^n}{\partial a_1 ... \partial a_n} F(a_1, ..., a_n)$$

Equivalence of Random Variables

We will often be interested of identifying properties defined on Ω that can be seen to hold with probability 1, or, as is the standard term, *almost surely*[36]:

Definition 31 (Almost Surely)**.** Fix a probability space $(\Omega, \mathfrak{F}, P)$. A property R of elements of Ω is said to hold *almost surely* iff

$$\{\omega \mid R(\omega) \text{ holds}\} \in \mathfrak{F} \quad \text{and} \quad P(\{\omega \mid R(\omega) \text{ holds}\}) = 1$$

An important application of this notion is the following. Let x, y be two random variables carried by $(\Omega, \mathfrak{F}, P)$. We wish to call them **equivalent** iff they are equal almost surely. To ensure this definition is valid, we need to show that the truth-set of this property is in $\mathfrak{F}$, for an arbitrary choice of space, x and y.

Proposition 32. *Let x, y be random variables carried by $(\Omega, \mathfrak{F}, P)$. Then,*

$$Z =_{df} \{\omega \mid x(\omega) = y(\omega)\} \in \mathfrak{F}$$

Proof. We will use the following lemma:

Lemma. *Let the function $y : \Omega \to \mathbb{R}$ be defined as follows:*

$$y(\xi) =_{df} f(x_1(\xi), ..., x_n(\xi))$$

Whenever $f : \mathbb{R}^n \to \mathbb{R}$ is a Borel *function, y is a random variable.*

Proof of Lemma. A composition of measurable functions is measurable. □

Consider now the probability function $w(\omega) =_{df} x(\omega) - y(\omega)$. Since $f(x, y) = x - y$ is a continuous function, it is also Borel (recall the Lemma in Section 2.2). Therefore w is a random variable, which implies that $w^{-1}(\{0\})$ is in $\mathfrak{F}$, since $\{0\}$ is a Borel set. This completes the proof, since

$$w^{-1}(\{0\}) =_{df} \{\omega \mid w(\omega) =_{df} x(\omega) - y(\omega) = 0\} = Z$$ □

We are now assured the following definition is valid:

Definition 32. Two random variables x and y carried by a probability space $(\Omega, \mathfrak{F}, P)$ are said to be *equivalent* iff they are equal almost surely.

We now show that any two equivalent random variables carried by a space $(\Omega, \mathfrak{F}, P)$ induce the *same* measure over $\mathbb{R}$ and therefore also have identical distribution functions. In this sense, any results we may prove about a certain random variable x only by appeal to its distribution or induced measure can at best identify x up to equivalence.

[36] In real analysis, the counterpart term is 'almost everywhere'. This is inappropriate here since it fails to emphasize the dependence on the particular choice of probability measure, which might not be the Lebesgue measure. Measure theorists often make this dependence explicit by writing "almost everywhere (μ)", where μ is the particular measure in question. We adopt Kolmogorov's 'almost surely' as a midway solution.

Proposition 33. *If x and y are equivalent, then*

$$\text{for all } A \in \mathcal{B}(\mathbb{R}), \quad P^{(x)}(A) = P^{(y)}(A) \tag{2.44}$$

Proof. Considering Z as before:

$$\begin{aligned} P(x^{-1}(A)) &= P(x^{-1}(A) \cap Z^c) + P(x^{-1}(A) \cap Z), \text{ by additivity} \\ &= P(x^{-1}(A) \cap Z^c), \text{ since } P(x^{-1}(A) \cap Z) \leq P(Z) = 0 \end{aligned}$$

Similarly, we get $P(y^{-1}(A)) = P(y^{-1}(A) \cap Z^c)$. But we now observe that:

$$\begin{aligned} x^{-1}(A) \cap Z^c &= \{\omega \mid x(\omega) = y(\omega) \in A\} = y^{-1}(A) \cap Z^c \\ \therefore \quad P(x^{-1}(A)) &= P(x^{-1}(A) \cap Z^c) = P(y^{-1}(A) \cap Z^c) = P(y^{-1}(A)). \end{aligned}$$

□

Convergence of Sequences of Random Variables

Consider now a sequence of random variables $x_1, x_2, \ldots$. We first prove that the proposition "$(x_i : i \in \mathbb{N})$ converges" corresponds to an event in $\mathfrak{F}$, which we call the *convergence set* of $(x_i : i \in \mathbb{N})$.

Proposition 34. $A =_{df} \{\xi \in \Omega \mid$ *the sequence* $x_1(\xi), x_2(\xi), \ldots$ *converges*$\} \in \mathfrak{F}$

Proof. We write down what this set A is by definition of (Cauchy) convergence:

$$\begin{aligned} \xi \in A &\Leftrightarrow \forall k \, \exists n \, \forall m, m' \geq n, |x_m(\xi) - x_{m'}(\xi)| \leq 1/k \\ \therefore \qquad A &= \bigcap_k \bigcup_n \bigcap_{m \geq n} \bigcap_{m' \geq n} \{\xi : |x_m(\xi) - x_{m'}(\xi)| \leq 1/k\} \end{aligned}$$

But, for each fixed m, m', k, the set $\{\xi : |x_m(\xi) - x_{m'}(\xi)| \leq 1/k\}$ is the inverse image of the Borel set $(-\infty, 1/k]$ under the function:

$$a(\xi) =_{df} |x_m(\xi) - x_{m'}(\xi)|$$

However, a is a random variable, since the function $f(x, y) =_{df} |x - y|$ is continuous, hence Borel. So $A \in \mathfrak{F}$ as required. □

We can now speak of the probability of convergence of a sequence of random variables. In fact, the limit of this sequence, if it exists[37], is itself a random variable.

Proposition 35. *The function defined below is a random variable:*

$$x(\xi) =_{df} \begin{cases} \lim_{n \to \infty} x_n(\xi), & \text{if } \xi \in A \\ 0, & \text{otherwise} \end{cases} \tag{2.45}$$

[37] In most modern textbooks, random variables are understood to be functions from Ω to the extended real field. Under this understanding, the qualification 'if the limit exists' above only serves to cover cases of oscillating divergence and can be dropped altogether in the case of lim sup's and lim inf's.

Proof. We need to show that the set $B =_{df} \{\xi \mid x(\xi) < a\}$ is in $\mathfrak{F}$ for each a. We take cases, depending on whether $a \leq 0$ or $a > 0$. In case $a \leq 0$, then:

$$\xi \in B \;\Rightarrow\; \text{the sequence } (x_i(\xi) : i \in \mathbb{N}) \text{ converged}$$

since otherwise, by convention, $x(\xi) = 0$ which contradicts $x(\xi) < a \leq 0$. So:

$$\begin{aligned}\{\xi \mid x(\xi) < a\} &= A \cap \{\xi \mid \text{the limit is strictly less than } a\} \\ &= A \cap \{\xi \mid \exists\, q \in \mathbb{Q}\; \exists N \in \mathbb{N}\; \forall\, n > N,\; x_n < a - q\} \\ &= A \cap \bigcup_{q \in \mathbb{Q}} \bigcup_{N \in \mathbb{N}} \bigcap_{n > N} \{\xi \mid x_n < a - q\} \in \mathfrak{F}\end{aligned}$$

This concludes the case $a \leq 0$. If $0 < a$, the event $x(\xi) < a$ can occur either because of convergence as above, or because the sequence in fact diverged, in which case $x(\xi)$ is set to $0 < a$ by convention. Hence:

$$\{\xi \mid x(\xi) < a\} = A^c \cup \left(A \cap \bigcup_{q \in \mathbb{Q}} \bigcup_{N \in \mathbb{N}} \bigcap_{n > N} \{\xi \mid x_n < a - q\} \right) \in \mathfrak{F} \qquad \square$$

It is useful to distinguish between the following three different types of convergence of sequences of random variables.

Definition 33. Recall that $(x_i : i \in \mathbb{N})$ is said to converge *pointwise* to the random variable x, given by (2.45), iff $A = \Omega$. We now say that $(x_i : i \in \mathbb{N})$ converges *almost surely* to the random variable x, given by (2.45), iff:

$$P(A) = 1 \tag{2.46}$$

Moreover, we say that $(x_i : i \in \mathbb{N})$ converges *in probability* to the random variable x, given by (2.45), iff, for every $\epsilon > 0$:

$$\lim_{n \to \infty} P(\{\omega : |x_n(\omega) - x(\omega)| > \epsilon\}) = 0 \tag{2.47}$$

Kolmogorov proves several results concerning convergence which we summarise in the following theorem:

Theorem 11. *Consider a sequence of random variables $(x_i : i \in \mathbb{N})$ and the respective sequence of distribution functions $(F_i : i \in \mathbb{N})$. Then:*

I. if $(x_i : i \in \mathbb{N})$ converges almost surely to x, then it also converges in probability to x

II. if $(x_i : i \in \mathbb{N})$ converges in probability to x and also to x', then x and x' are equivalent

III. if $(x_i : i \in \mathbb{N})$ converges in probability to x, then $(F_i : i \in \mathbb{N})$ converges to the distribution function F of x at each point of continuity of F.

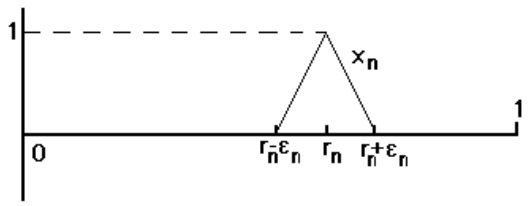

Figure 2.1: A graphical illustration of the definition of $x_n : \mathbb{R} \to \mathbb{R}$

Proof. We only prove (I) here, since the rest of the proofs are as direct and can be found in the Grundbegriffe. Assume $(x_i : i \in \mathbb{N})$ converges a.s. to x and let A be the convergence set of $(x_i : i \in \mathbb{N})$. Now pick $\epsilon > 0$. Then for every $\xi \in A$,

$$\exists\, n\, \forall\, p\ \ |x_{n+p}(\xi) - x(\xi)| < \epsilon$$

$$\therefore \quad A \subseteq \bigcup_n S_n \tag{2.48}$$

where $S_n =_{df} \bigcap_p \{\xi \mid |x_{n+p}(\xi) - x(\xi)| < \epsilon\}$. We now observe that

$$m' \geq m \ \Rightarrow\ S_{m'} \supseteq S_m \tag{S1}$$

$$\text{and}\ \ S_n \subseteq \{\xi : |x_n(\xi) - x(\xi)| < \epsilon\} \tag{S2}$$

Therefore,

$$\begin{aligned} P(A) &\leq \bigcup_n S_n, \text{ by (2.48)} \\ &= \lim_{n\to\infty} P(S_n), \text{ by (S1) and Axiom VI} \\ &\leq \lim_{n\to\infty} P(\{\xi : |x_n(\xi) - x(\xi)| < \epsilon\}), \text{ by (S2)} \end{aligned}$$

So, as required,

$$1 = P(A) \leq \lim_{n\to\infty} P(\{\xi : |x_n(\xi) - x(\xi)| < \epsilon\}) \leq 1 \qquad \square$$

We conclude this section by offering a counterexample to the converse of the statement proved above - a sequence $(x_i : i \in \mathbb{N})$ that converges in probability but fails to converge almost surely.

Proposition 36. *Convergence in probability $\not\Rightarrow$ Convergence almost surely.*

Proof. We will provide a counterexample in the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), Leb)$. Enumerate $\mathbb{Q} = \{r_1, r_2, ...\}$. Then define for each n, the random variable $x_n(\xi) : \mathbb{R} \to \mathbb{R}$ as indicated in Figure (2.1), where:

$$\epsilon_n = \frac{1}{2^{2n+1}}$$

In other words x_n is zero everywhere except in the region $(r_n - \epsilon_n, r_n + \epsilon_n)$, where it is linearly increasing in $(r_n - \epsilon_n, r_n]$ and linearly decreasing in $[r_n, r_n + \epsilon_n)$, attaining its maximum value 1 at r_n. Moreover each x_n is continuous, hence a random variable.

We now remark that for any $\epsilon > 0$ and letting x be the constant function 0:

$$\begin{aligned} P(\{\xi : |x_n(\xi) - x(\xi)| > \epsilon\}) &= Leb(\{\xi : |x_n(\xi) - x(\xi)| > \epsilon\}) \\ &\leq Leb(\{\xi : |x_n(\xi) - x(\xi)| > 0\}) = 2 \cdot \epsilon_n = \frac{1}{2^{2n}} \end{aligned}$$

So $(x_n : n \in \mathbb{N})$ converges in probability to the constant function 0:

$$\lim_{n\to\infty} P(\{\xi : |x_n(\xi) - x(\xi)| > \epsilon\}) \leq \lim_{n\to\infty} \frac{1}{2^{2n}} = 0$$

We will now show that convergence almost surely fails. The key observation is that, for any enumeration $(r_n)_n$ of the rationals, any of its tails, $(r_n)_{n>N}$, is *dense* in $[0, 1]$. This means that for any $\xi \in [0, 1]$, $(r_n)_n$ will return arbitrarily close to ξ infinitely often:

$$\forall\, \delta > 0\, \forall\, N\, \exists\, n > N\, , |r_m - \xi| < \delta \tag{2.49}$$

So now pick an arbitrary $\xi \in \mathbb{R}$ and let M be any number. If we set $N := M$ and $\delta := 0.5 \cdot \epsilon_M$ in (2.49), we obtain that

$$\exists n > M, \text{ for which } |r_n - \xi| < 0.5 \cdot \epsilon_M, \text{ hence } x_n(\xi) > 0.5.$$

With the same argument applied to $\xi' = \xi + 10$ we obtain that

$$\exists n' > M, \text{ for which } |r_{n'} - \xi| > \epsilon_M, \text{ hence } x_n(\xi) = 0.$$

Since for each M we can find such numbers n and n', we can construct two subsequences of $(x_n(\xi) : n \in \mathbb{N})$, one of which is bounded below by 0.5 and the other of which is identically zero. This implies that $(x_n(\xi) : n \in \mathbb{N})$ diverges.

Since ξ was arbitrary, we have just shown that the convergence set of the sequence of random variables $(x_n : n \in \mathbb{N})$ is empty and hence has probability 0 by the properties of measure. In particular, its probability is not 1, which completes the proof. □

2.5 Mathematical Expectations

We dedicate this section to the study of *mathematical expectations*, an application of Lebesgue integration on probability fields. Besides being very useful in applications of probability theory (by way of their intuitive meaning), expectations are invaluable as a formal tool for the study of random variables. In particular, they will play a central role in the definition of conditional probability that appears in the next section.

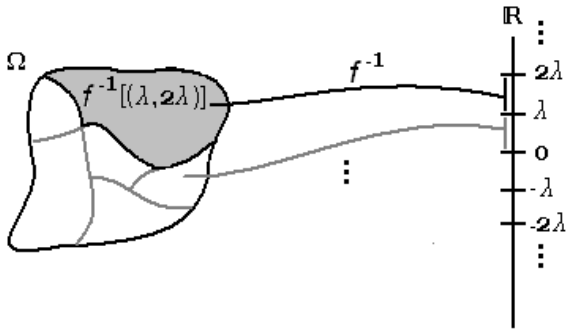

Figure 2.2: A graphical illustration of *Lebesgue's ladder*; the real line, which is the *image space* of f, is partitioned into intervals[39] of length λ. The inverse image operator f^{-1} then induces a partition over the *domain*, which is Ω.

Definition and Properties

Consider a probability space $(\Omega, \mathfrak{F}, P)$ equipped with a random variable $x : \Omega \to \mathbb{R}$ and A an arbitrary set of $\mathfrak{F}$. In a familiar approach, we will attempt to approximate the integral of x over A by weighted sums of the form $\sum_{B \in \mathfrak{U}} P(B) x_B$, where $\mathfrak{U}$ is some partition of A and x_B is some indicative value of the values $x(\omega)$ takes for $\omega \in B$. In particular, we consider the following series:

$$S_\lambda(x, A, P) =_{df} \sum_{k=-\infty}^{k=+\infty} k\lambda P(\{\omega \mid k\lambda \leq x(\omega) < (k+1)\lambda\} \cap A) \tag{2.50}$$

Remark. The partition $\mathfrak{U}$ that generates this series is not defined directly in the *domain* of x. Instead, we consider a regular partition of the *image space* of x, the real line, into intervals $[k\lambda, (k+1)\lambda)$ for $k \in \mathbb{Z}$. We then produce the partition $\mathfrak{U}$ by considering the inverse images of these intervals. In this manner, we exploit the measurability of x to ensure that our partition contains measurable sets only (ie $\mathfrak{U} \subseteq \mathfrak{F}$) and are also assured that the values that x takes on each interval $[k\lambda, (k+1)\lambda)$ are very close to our choice of indicative value $k\lambda$. This idea of carving up in a regular manner the image space as opposed to the domain is what differentiates Lebesgue integration from Riemann integration and has come to be known as *Lebesgue's Ladder* (see Figure 2.2).

In case the series $S_\lambda(x, A, P)$ converges absolutely for every λ and its limit as $\lambda \to 0$ exists, it is then defined to be the *Lebesgue integral* or simply *integral*[40] of x over A, relative to the probability measure P:

[40]The standard measure theoretic notation for this integral is $\int_A x dP$, which can however prove comfusing when x is a function of several variables or defined via composition. The notation in (2.51) makes the domain somewhat more explicit and is also closer to Kolmogorov's choice, $\int_A x P(dE)$.

Definition 34.

$$\int_A x(\omega)d_P(\omega) =_{df} \lim_{\lambda \to 0} S_\lambda \tag{2.51}$$

Definition 35. We call x *integrable on* A iff (2.51) exists.

The expectation of x can now be defined as the integral of x over the entire sample space Ω:

Definition 36. The *mathematical expectation* of an r.v. x is given by

$$E(x) =_{df} \int_\Omega x(\omega)d_P(\omega) \tag{2.52}$$

We now introduce certain useful properties of integrals and mathematical expectations in the following two theorems, which we state without proof:

Theorem 12. *The following all hold:*

I. If x is integrable on A, then it is integrable on any $A' \subseteq A$, $A' \in \mathfrak{F}$.

II. If x is integrable on $A = \bigcup_{n\in\mathbb{N}} A_n$, A_n pairwise disjoint sets of $\mathfrak{F}$, then:

$$\int_A x(\omega)d_P(\omega) = \sum_n \int_{A_n} x(\omega)d_P(\omega)$$

III. If x is integrable on A, then $|x|$ is also integrable on A and:

$$\left|\int_A x(\omega)d_P(\omega)\right| = \int_A |x(\omega)|\, d_P(\omega)$$

IV. If $\forall\omega \in A$, $0 \leq y(\omega) \leq x(\omega)$ and x is integrable on A, then y is also integrable on A and:

$$\int_A y(\omega)d_P(\omega) \leq \int_A x(\omega)d_P(\omega)$$

V. If $\forall\omega \in A$, $m \leq x(\omega) \leq M$, then:

$$mP(A) \leq \int_A x(\omega)d_P(\omega) \leq MP(A)$$

VI. If x and y are integrable on A, then $Kx + Ly$ is also integrable and:

$$\int_A (Kx(\omega) + Ly(\omega))d_P(\omega) = K\int_A x(\omega)d_P(\omega) + L\int_A y(\omega)d_P(\omega)$$

VII. Let $(x_n : n \in \mathbb{N})$ be a sequence of random variables such that

$$\sum_n \int_C |x_n(\omega)| \, d_P(\omega) < \infty$$

Then the random variable x defined by $x =_{df} \sum_n x_n$ converges pointwise on $C \setminus B$ for some set B such that $P(B) = 0$. Moreover, setting $x(\xi) = 0$ for $\xi \in B$, we get:

$$\int_A x(\omega) d_P(\omega) = \sum_n \int_A x_n(\omega) d_P(\omega)$$

VIII. If x and y are equivalent then for any $A \in \mathfrak{F}$:

$$\int_A x(\omega) d_P(\omega) = \int_A y(\omega) d_P(\omega) \tag{2.53}$$

IX. If (2.53) holds for every $A \in \mathfrak{F}$, then x, y are equivalent.

X. Every bounded random variable is integrable.

Proof. Not included. Refer to [9] and [8]. □

Remark. Property IX makes it possible to implicitly specify a function almost everywhere by explicitly specifying all its integrals. This will be crucial in the construction of conditional probability with respect to probability 0 events in the next subsection.

The following properties of mathematical expectations now follow directly from its definition and Theorem 12:

Theorem 13. *The following hold:*

I. $|E(x)| \leq E(|x|)$

II. If $0 \leq y \leq x$ everywhere, then $E(y) \leq E(x)$.

III. $\inf(x) \leq E(x) \leq \sup(x)$

IV. $E(Kx + Ly) = KE(x) + LE(y)$

V. $E(\sum_n x_n) = \sum_n E(x_n)$ *if* $\sum_n E(|x_n|)$ *converges.*

VI. If x,y are equivalent, then $E(x) = E(y)$.

VII. Every bounded random variable has a mathematical expectation.

Proof. Trivially follows from Theorem 12. □

Elementary Conditional Mathematical Expectations

It is useful to introduce a concept of *elementary conditional mathematical expectation*, by employing the elementary conditional probability measure P_B which we encountered in Section 2.1:

Definition 37. Let x be a random variable and let $B \in \mathfrak{F}$, with $P(B) > 0$. The *conditional mathematical expectation of x with respect to B* is given by:

$$E_B(x) = \int_\Omega x(\omega) d_{P_B}(\omega) \tag{2.54}$$

where P_B is defined as in (2.1) and is a probability measure by Proposition 4:

$$P_B(A) =_{df} \frac{P(A \cap B)}{P(B)}, \quad (A \in \mathfrak{F})$$

Conditional expectations are most useful when written in a different form, which we now prove equivalent to (2.54):

Proposition 37. *For $P(B) \neq 0$:*

$$E_B(x) = \frac{1}{P(B)} \int_B x(\omega) d_P(\omega) \tag{2.55}$$

In particular, we get that:

$$E(x) = P(B)E_B(x) + P(B^c)E_{B^c}(x) \tag{2.56}$$

Proof. We observe that:

$$E_B(x) =_{df} \int_E x P_B(dE) = \int_B x(\omega) d_{P_B}(\omega) + \int_{B^c} x(\omega) d_{P_B}(\omega)$$

by Theorem 12.II. Since P_B is 0 everywhere on B^c, the series $S_\lambda(x, B^c, P_B)$ consists of zero terms only and respectively the limit as $\lambda \to 0$ must also be zero, so:

$$\int_{B^c} x(\omega) d_{P_B}(\omega) = 0$$

On the other hand, we observe that only sets intersected with B are featured in the series $S_\lambda(x, B, P)$. Therefore, conditioning with respect to B has the following effect:

$$S_\lambda(x, B, P_B) = \frac{1}{P(B)} S_\lambda(x, B, P)$$

$$\therefore \quad \int_B x(\omega) d_{P_B}(\omega) = \frac{1}{P(B)} \int_B x(\omega) d_P(\omega)$$

as required for (2.55). Now, (2.56) follows easily from (2.55):

$$\begin{aligned} E_B(x) &= \frac{1}{P(B)} \int_B x(\omega) d_P(\omega), \text{ by (2.55)} \\ &= \frac{1}{P(B)} \left(\int_\Omega x(\omega) d_P(\omega) - \int_{B^c} x(\omega) d_P(\omega) \right), \text{ by Theorem 12.(II)} \\ &= \frac{1}{P(B)} \left(E(x) - E_{B^c}(x) \cdot P(B^c) \right), \text{ again by (2.55)} \end{aligned}$$ □

Differentiation and Integration of Expectations

Although we will not need the results of this section for the purpose of defining conditional probabilities, we provide them along with an example of their application as further indication of the power of Kolmogorov's formalism in dealing with abstract questions in geometric probability.

Fix a probability space $(\Omega, \mathfrak{F}, P)$ and consider a function $f : \Omega \times \mathbb{R}^n \to \mathbb{R}$,

$$f(\omega, t) \in \mathbb{R}, \quad (\omega \in \Omega,\ t \in \mathbb{R}^n)$$

Let us call such a function an Ω-*functional*. It can be viewed as a family of real functions indexed by $\omega \in \Omega$ or equivalently as a family of probability functions indexed by $t \in \mathbb{R}^n$. Accordingly, an Ω-functional can be acted upon in various ways, in particular:

- By evaluation of its first argument at $\omega \in \Omega$, which returns a real function $\mathbb{R} \to \mathbb{R}$, denoted by $f_\omega(t)$.
- By evaluation of its second argument, which returns a probability function $\Omega \to \mathbb{R}$ denoted by $f_t(\omega)$.
- By differentiation with respect to its second argument, which (when all derivatives exist) returns an Ω-functional $(\Omega \times \mathbb{R}) \to \mathbb{R}$, denoted by:

$$\frac{\partial}{\partial t} f(\omega, t)$$

- By Lebesgue integration along its first argument over a set $A \in \mathfrak{F}$, which (when all integrals exist) returns a real function $\mathbb{R} \to \mathbb{R}$, denoted by:

$$\int_A f(\omega, t) d_P(\omega)$$

  Remark. As a special case, taking the expectation of f along its first argument falls under this type of operation.

- By Riemann integration along its second argument over a rectangle $S = \prod_{i=1}^n [a_i, b_i]$, which (when all integrals exist) returns a probability function $\Omega \to \mathbb{R}$, denoted by:

$$\int_S f(\omega, t) dt =_{df} \int_{a_1}^{b_1} \int_{a_2}^{b_2} ... \int_{a_n}^{b_n} f(\omega, t) dt_1 dt_2 ... dt_n$$

Observe, however that the series $S_\lambda(x, A, P)$ via which we defined the Lebesgue integral can only be defined for x a random variable (since the inverse images of all intervals must be in $\mathfrak{F}$). This makes the following definition natural:

Definition 38. If x is an Ω-functional and for all t, $x_t(\xi)$ is a random variable, then let us call x a *random Ω-functional*.

Finally, we introduce some useful notation for the taking the expectation of f along its first argument (provided the Lebesgue integral converges):

Definition 39. If f is a random Ω-functional, then we define

$$E(f)(t) =_{df} \int_\Omega f(\omega, t) d_P(\omega), \text{ if it exists}$$

It is customary upon encountering a function on the cartesian product of two distinct spaces to enquire whether a change of the order of certain operations is allowed. We will now demonstrate two cases in which such a change is allowed. Both results can be found in Section IV, Paragraph 5 of the Grundbegriffe, with complete proofs, although Kolmogorov's notation makes the proofs somewhat hard to follow. We omit the proof of Theorem 14 since we will not be using this Theorem again. However, we do reproduce the proof of Theorem 15, with different notation and some gaps filled, since this theorem will serve as the basis for some discussion later.

Theorem 14 (Leibniz's Rule)**.** *Let $x(\omega, t)$ be a random Ω-functional, where $\omega \in \Omega$ and $t \in \mathbb{R}$ (ie we take $n = 1$ in the original definition). Now assume that the following conditions hold:*

1. *for each t, $E(f)(t)$ exists*
2. *for all ω, $f_\omega(t)$ is an everywhere differentiable real function*
3. *there exists a single constant M, such that for all ω, $\frac{\partial}{\partial t} f(\omega, t)$ is bounded in absolute value by M.*

Then $\frac{\partial}{\partial t} f(\omega, t)$ is also a random Ω-functional and:

$$\frac{d}{dt} E(f)(t) = E\left(\frac{\partial}{\partial t} f(\omega, t)\right) \tag{2.57}$$

Proof. Omitted. Can be found in [6, p.44]. □

Theorem 15. *Let f be a random Ω-functional. If f is bounded in absolute value by some constant K and each real function f_ω is integrable in the Riemann sense, then:*

$$\int_a^b f(\omega, t) dt$$

is a random variable. Moreover, its expectation exists and is equal to:

$$\int_a^b E(f)(t) dt = E(\int_a^b f(\omega, t) dt) \tag{2.58}$$

Proof. For any ω, the real function f_ω is Riemann integrable, so:

$$J(\omega) =_{df} \int_a^b f(\omega, t)dt$$

exists, which of course yields a probability function $J : \omega \mapsto J(\omega)$. Now define:

$$S_n(\omega) =_{df} \frac{1}{h}\sum_{k=1}^{n} f_{a+kh}(\omega), \text{ where } h =_{df} \frac{b-a}{n} \tag{2.59}$$

Then for each n, S_n is a random variable, since it is a scalar multiple of a sum of random variables (the f_{a+kh}'s). Moreover, for any fixed ω, the sequence $(S_n(\omega) : n \in \mathbb{N})$ satisfies:

$$\lim_{n\to\infty} S_n(\omega) \text{ exists and is equal to } J(\omega)$$

by definition of the Riemann integral.

Clearly then the sequence $(S_n : n \in \mathbb{N})$ of random variables converges pointwise to the probability function J. Pointwise convergence is a (trivial) instance of convergence almost surely, which implies by Proposition 35 that J is a random variable, too. Moreover, $J(\omega) \leq K(b-a)$, by (2.59), since f is bounded above by K. So J is in fact a *bounded* random variable, which means its expectation $E(J)$ exists by VII of Theorem 13.

Convergence a.s. also implies that the S_n's converge to J in probability, by I of Theorem 13. This is to say that

$$\forall \epsilon > 0,\ \lim_{n\to\infty} P(\{\omega : |S_n(\omega) - J(\omega)| > \epsilon\}) = 0$$

which in particular implies

$$\forall \epsilon > 0, \exists N, \forall n \geq N,\ P(\{\omega : |S_n(\omega) - J(\omega)| > \epsilon\}) < \epsilon \tag{2.60}$$

For readability in what follows, we denote the set involved in (2.60) as follows:

$$A =_{df} \{\omega : |S_n(\omega) - J(\omega)| > \epsilon\}) \tag{2.61}$$

Now for each n, S_n is a random variable bounded above by $K(b-a)$, so its expectation exists and is equal to:

$$E(S_n) = \frac{1}{h}\sum_{k=1}^{n} E(f_{t+kh}), \text{ by (IV) of Theorem 13.}$$

We finally have:

$$\begin{aligned}
&|E(S_n) - E(J)| = |E(S_n - J)|\text{, by Theorem 13.(IV)}\\
&\leq E(|S_n - J|), \text{ by 13.(I)}\\
&= P(A)E_A(|S_n - J|) + P(A^c)E_{A^c}(|S_n - J|), \text{ by (2.56)}\\
&\leq P(A)E_A(|S_n| + |J|) + P(A^c)E_{A^c}(|S_n - J|), \text{ by 13.(II) and triangle inequality}\\
&\leq P(A)(E_A(|S_n|) + E_A(|J|)) + P(A^c)E_{A^c}(|S_n - J|), \text{ by 13.(IV)}\\
&\leq 2K(b-a)P(A) + P(A^c)E_{A^c}(|S_n - J|)
\end{aligned} \tag{2.62}$$

by (III) of Theorem 13 and since both $|J|$ and $|S_n|$ are bounded by $K(b-a)$.

We now observe that we can bound strictly the rightmost summand in (2.62) by ϵ. If $P(A^c) = 0$, then trivially the rightmost summand is $0 < \epsilon$. If not, by definition of A, we have that for all $\omega \notin A$, $|S_n(\omega) - J(\omega)| \leq \epsilon$. Therefore:

$$\begin{aligned} & \sup\{|S_n(\omega) - J(\omega)| : \omega \in A^c\} \leq \epsilon \\ \therefore \quad & E_{A^c}(|S_n(\omega) - J(\omega)|) \leq \epsilon, \text{ by 13.(III)} \\ \therefore \quad & P(A^c)E_{A^c}(|S_n(\omega) - J(\omega)|) \leq \epsilon, \text{ since } P(A^c) \leq 1. \end{aligned}$$

Overall then we have:

$$\begin{aligned} |E(S_n) - E(J)| &\leq 2K(b-a)P(A) + \epsilon \\ &\leq (2K(b-a) + 1)\epsilon, \text{ by (2.60)}. \end{aligned}$$

Therefore

$$\begin{aligned} & \lim_{n\to\infty} E(S_n) = E(J) \\ \therefore \quad & \int_a^b E(f)(t)dt = E(J) \end{aligned}$$

assuming $E(f)(t)$ is Riemann integrable.

Remark. Observe that the existence of $\int_a^b E(f)(t)dt$ implies the existence of $\lim_n E(S_n)$ but not vice versa, which is why we need to separately assume that $E(f)(t)$ is Riemann integrable. □

The above proof holds as is for a double integral, that is, where f is now a family of random variables indexed by $(x, y) \in \mathbb{R}^2$:

Theorem 16. *Let f be a family of random variables indexed by $(x, y) \in \mathbb{R}^2$. Then it is also a family of functions $f_\omega : \mathbb{R}^2 \to \mathbb{R}$ indexed by $\omega \in \Omega$. If all these latter are bounded in absolute value by some constant K and are integrable in the Riemann (double integral) sense, and the expectation $E(f)(x, y)$ is also Riemann integrable, then $\int_a^b \int_c^d f(\omega, x, y)\, dxdy$ is a random variable and:*

$$\int_a^b \int_c^d E(f)(x, y)\, dxdy = E\left(\int_a^b \int_c^d f(\omega, x, y)\, dxdy\right) \tag{2.63}$$

Change of order of integration: Theorem 15 or Fubini's Theorem?

Observe that Theorems 15 and 16 are results that guarantee a change in the order of integration is possible (although one of the two integrals is Lebesgue and the other Riemann). It is then natural to attempt a comparison between this theorem and the standard measure-theoretic result that is customarily used to justify a change in the order of integration, *Fubini's Theorem*:

Theorem (Fubini's Theorem)**.** *Consider two probability spaces* $(\Omega_1, \mathfrak{F}_1, \mu_1)$*,* $(\Omega_2, \mathfrak{F}_2, \mu_2)$ *and consider a* bounded random variable *defined on the cartesian product* $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$*, equipped with the product* σ*-algebra* $\mathfrak{F}_1 \times \mathfrak{F}_2$*. Now consider the following two probability functions:*

$$I_1^f(\omega_1) =_{df} \int_{\Omega_2} f(\omega_1, \omega_2) d_{\mu_2}(\omega_2), \qquad I_2^f(\omega_2) =_{df} \int_{\Omega_1} f(\omega_1, \omega_2) d_{\mu_1}(\omega_1)$$

Then both I_1^f *and* I_2^f *are bounded random variables and:*

$$\int_{\Omega_1} I_1^f(\omega_1) d_{\mu_1}(\omega_1) = \int_{\Omega_2} I_2^f(\omega_2) d_{\mu_2}(\omega_2)$$

Proof. Omitted, can be found in [15]. □

In the next subsection, we attempt to compare the applicability of the two theorems by way of an example in geometric probability.

An Example: Splatters on a Wall

The prototypical experiment in geometric probability on the plane is one of 'picking a point at random' from a bounded region S of $\mathbb{R}^2$, where without loss of generality we can assume $S = [a, b] \times [c, d]$). This is usually modelled by the probability space $(S, \mathcal{B}(S), Leb)$, or, more generally, $(S, \mathcal{B}(S), P)$.

Kolmogorov invites us to consider a different kind of experiment where we are picking *regions* ('shapes') at random, not points. One is invited to throw paint splatters on a wall, so to speak. We are hence looking to model this experiment by a probability space of the form $(\Omega, \mathfrak{F}, \mu)$, where Ω is the family of allowable splatters. Certainly it is natural to assume that splatters be Borel regions:

$$\Omega \subseteq \mathcal{B}(S)$$

Note then that μ assigns probabilities *not* to individual spatters, but rather to *families of splatters.*

An Interesting Duality

Consider now the following statement:

$$\text{"splatter } G \text{ turns out to contain } (x, y)\text{"} \tag{S1}$$

We are accustomed to contexts where it is the point (x, y) that is randomly selected, whence the probability of (S1) is given by $P(G)$. However, in the context of the probability space $(\Omega, \mathfrak{F}, \mu)$, one imagines a fixed point on the wall and considers the probability that it ends up painted (ends up in the region randomly selected):

$$Pr(x, y) =_{df} \mu(\{G \in \Omega \mid (x, y) \in G\}) \tag{2.64}$$

This duality is perhaps best seen by examining the indicator function of (S1):

$$f(G,(x,y)) =_{df} \begin{cases} 1, & \text{if } (x,y) \in G \\ 0, & \text{otherwise} \end{cases} \qquad (G \in \mathcal{B}(S),\ (x,y) \in S) \tag{2.65}$$

For a fixed choice of G, f_G is a probability function carried by the space $(S, \mathcal{B}(S), Leb)$, whereas for fixed (x,y), $f_{(x,y)}$ is a probability function carried by the space $(\mathcal{B}(S), \mathfrak{F}, P')$.

Remark. Note that for the above argument to go through, $Pr(x,y)$ as defined in (2.64) must be the probability of a set in $\mathfrak{F}$, otherwise it does not exist. Therefore, the following assumption must be made:

$$X =_{df} \{G \in \Omega \mid (x,y) \in G\} \in \mathfrak{F} \tag{A1}$$

We will now investigate the probability space $(\Omega, \mathfrak{F}, \mu)$. It will turn out that our initial assumption that $\Omega \subseteq \mathcal{B}(S)$ is too weak to allow us to argue with Riemann integrals, we therefore restrict further our set of elementary events:

$$\Omega \subseteq \mathcal{RI}(S), \quad \text{ie} \quad f_G(x,y) \text{ is Riemann integrable}$$

An interesting question that we may ask of our space is "what is the expected area of a randomly selected splatter?" The following recipe for the computation of this value seems intuitively appealing: weigh the presence of each (x,y) by the probability $Pr(x,y)$ that it should belong to a randomly selected splatter - integrate over S. However, the formal response involves instead a much more complicated integral over all splatters, rather than one over all points. The comparison of these two approaches gives us the opportunity we were looking for so as to contrast Theorem 16 with Fubini's Theorem.

Theorem 17. *Assume that*[41]*:*

$$X =_{df} \{G \in \mathcal{B}(S) \mid (x,y) \in G\} \in \mathfrak{F}; \tag{A1}$$

$$Leb \text{ is a random variable as a probability function on } \Omega; \tag{A2}$$

$$Pr(x,y) \text{ is integrable as a function of } (x,y) \text{ in the Riemann sense;} \tag{A3}$$

then

$$\int_a^b \int_c^d Pr(x,y)dxdy = E(Leb) \tag{2.66}$$

Proof. First, we need a lemma:

Lemma. *For each fixed (x,y), the function $f_{(x,y)}$ is a random variable and:*

$$Pr(x,y) = E(f_{(x,y)}) \tag{2.67}$$

[41] The extra assumptions A3, A4 are in fact the minimal assumptions such that the Riemann integral on the LHS of (2.66) and the Lebesgue integral on its RHS be well-defined.

Proof of Lemma. Observe that $f_{(x,y)}$ is the indicator function of X. Hence:

$$f^{-1}_{(x,y)}(-\infty, a) =_{df} \{G \mid f_{(x,y)}(G) < a\} = \begin{cases} \Omega, & \text{if } a \geq 1 \\ X^c, & \text{if } 0 \leq a < 1 \\ \emptyset, & \text{if } a < 0 \end{cases} \tag{2.68}$$

Clearly $\Omega, \emptyset \in \mathfrak{F}$, but also $X^c \in \mathfrak{F}$ by (A1). Also clearly:

$$E(f_{(x,y)}) = \mu(X) =_{df} Pr(x, y)$$

by the properties of the integral when applied to indicator functions. □

Observe now that, since $G \in \mathcal{RI}(S)$, the Lebesgue integral of the indicator function of G matches its Riemann integral (because whenever the Riemann and the Lebesgue integrals both exist, they agree - see Appendix on Integration):

$$Leb(G) =_{df} \int_S f_G(x, y) d_{Leb}(x, y) = \int_a^b \int_c^d f_G(x, y) dx\, dy$$

Using this to write out fully either side of (2.66) we get:

$$RHS = E(Leb) =_{df} \int_\Omega Leb(G) d_\mu(G) = \int_\Omega \int_a^b \int_c^d f_G(x, y)\, dx\, dy\, d_\mu(G)$$

$$LHS = \int_a^b \int_c^d Pr(x, y) dx dy = \int_a^b \int_c^d \int_\Omega f_{(x,y)}(G) d_\mu(G)\, dx\, dy$$

This makes (2.66) a straightforward application of Theorem 16, provided we establish that f satisfies its requirements. Indeed, by the Lemma, f is a bounded random Ω-functional and by assumptions (A2) and (A3), f and $E(f)$ are both Riemann integrable as functions of (x, y). □

We now observe that the same theorem can be inferred under different assumptions using Fubini's Theorem:

Theorem 18. *Consider the product σ-algebra $\mathfrak{F} \times \mathcal{B}(S)$. Now define D to be the truth-set of the relation $(x, y) \in G$:*

$$D =_{df} \{(G, (x, y)) \in \Omega \times S \mid (x, y) \in G\} = \bigcup_{G \in \Omega} (\{G\} \times G)\,.$$

Now assume that

$$D \in \mathfrak{F} \times \mathcal{B}(S). \tag{A4}$$

Then the following holds, where the LHS integral is Lebesgue:

$$\int_S Pr(x, y) d_{Leb}(x, y) = E(Leb). \tag{2.69}$$

Proof. We will apply Fubini on the spaces $(S, \mathcal{B}(S), Leb)$, $(\Omega, \mathfrak{F}, \mu)$ and the function f. We need only show that f is measurable in the product sense, but this holds precisely iff $D \in \mathfrak{F}$. The result then follows from (A4) and the observation:

$$Leb(G) = \int_S f_G(x,y) d_{Leb}(x,y) \qquad \square$$

Remark. That the proof we have just provided is also a proof of Theorem 17 under the extended set of assumptions (A1)-(A4) follows from the observation that, if $Pr(x,y)$ is Riemann integrable, the integral on the LHS of (2.69) can be replaced by a Riemann integral, hence yielding (2.66).

We pause to make explicit the simple relationship that exists between the set of assumptions (A1)-(A3) and the set of assumptions (A1)-(A4):

Proposition 38. *Assumptions (A1) and (A2) hold precisely whenever f is separately measurable with respect to each argument. Therefore:*

$$(A4) \Rightarrow (A1) \text{ and } (A2)$$

Proof. Assume that f is measurable in the product sense. Then, from standard results, it is also measurable separately, so:

$$X =_{df} f^{-1}_{(x,y)}(\{1\}) \in \mathfrak{F}$$

This yields (A1). Now consider the quantity $I^f_1(\omega_1)$ that appears in the statement of Fubini's Theorem. By setting $\Omega_1 =_{df} \Omega$ and $\Omega_2 =_{df} S$ (and substituting in the respective σ-algebras and measures), we obtain that:

$$I^f_1(G) =_{df} \int_S f(G,(x,y)) d_{Leb}(x,y) = Leb(G)$$

It forms part of the conclusion that I^f_1 is measurable, which yields (A3). $\square$

What is interesting here is that, despite appearances, none of the two theorems can be seen as a stronger version of the other. Theorem 18 does yield Theorem 17 as a special case, but rests on a stronger hypothesis, A4. Conversely, Theorem 17 holds under a *weaker* set of hypotheses, which makes it a stronger result, but it is also of more limited applicability, since it applies only to $\Omega \subseteq \mathcal{RI}(S)$, whereas Theorem 18 can apply to any $\Omega \subseteq \mathcal{B}(S)$. So there is a genuine question here as to which approach is better suited to model our intuitive understanding of what "picking regions at random" means. In what follows, we will provide some insight into this question by way of examples, although we will unfortunately fail to settle it.

Various Ways of Throwing Splatters - the countable case

We will now produce specific examples of spaces $(\Omega, \mathfrak{F}, \mu)$, where $\Omega \subseteq \mathcal{RI}(S)$, and investigate whether it is more natural to invoke Theorem 17 together with

the set of assumptions (A1)-(A3) or Theorem 18 together with the single assumption (A4).

We first show that in the countable case the question is trivialised.

Example 1 (The Countable Case)**.** *Let $\Omega = \{S_1, S_2, ...\}$ be any countable family of Riemann Integrable regions and $\mathfrak{F}$ be a σ-algebra over Ω that contains the singletons. Then let μ be the (unique) measure generated by the following values on the singletons:*

$$\mu(\{S_i\}) =_{df} p_i$$

Then all of (A1)-(A4) hold of $(\Omega, \mathfrak{F}, \mu)$.

Proof. Recall that D as expressed in (A4) is given by

$$D = \bigcup_{G \in \Omega} \{G\} \times G$$

Since the singletons $\{G\}$ are in $\mathfrak{F}$ and $G \in \Omega \subseteq \mathcal{B}(S)$, the expression above is a countable union of elements of $\mathfrak{F} \times \mathcal{B}(S)$ and so (A4) holds; hence also (A1) and (A2) hold.

To prove that (A3) also holds, we first observe that, by countable additivity:

$$Pr(x, y) =_{df} \mu(\{S_i \mid (x, y) \in S_i\}) = \sum_i f_{S_i}(x, y) p_i$$

We now need a lemma:

Lemma. *The following limit converges uniformly with respect to x and y:*

$$\lim_{n \to \infty} \sum_{i=1}^{n} f_{S_i}(x, y) p_i = Pr(x, y)$$

Proof of Lemma. We know that, for all (x, y),

$$\sum_{i=1}^{m} f_{S_i}(x, y) p_i \leq \sum_{i=1}^{m} p_i$$

since $0 \leq f_{S_i} \leq 1$. Then, as the partial sums $\sum_{i=1}^{m} p_i$ converge, so do their weighted versions, $\sum_{i=1}^{m} f_{S_i}(x, y) p_i$, which proves that the convergence rate is independent of (x, y). □

We now invoke two standard results: firstly, any finite linear combination of Riemann integrable functions is also Riemann integrable. This makes each partial sum a Riemann integrable function, since it is a linear combination of finitely many indicator functions of Riemann Integrable regions. Secondly, if $(f_n : n \in \mathbb{N})$ is a sequence of Riemann integrable functions that converge uniformly to f, then f is Riemann integrable. This result applied to the sequence of partial sums implies that $Pr(x, y)$ is Riemann Integrable, yielding (A3). □

Various Ways of Throwing Splatters - the uncountable case

We will now move to the uncountable case and begin by providing a general framework which captures all cases where Ω can be parameterised by n real parameters. Riemann Integrability here becomes quite hard to deal with in general, so we discuss it individually for each example.

Proposition 39 (A Framework for the Uncountable Case)**.** *Let our set of elementary events be parameterised by n real parameters $x_1, ..., x_n$ via an injective function ϕ, ie:*

$$\Omega =_{df} \{G_{\vec{z}} \mid \vec{z} \in [0,1]^n\} \quad \text{and} \quad \phi(G_{\vec{z}}) = \vec{z}$$

Now define $\mathfrak{F}$ via the the Borel sets in $[0,1]^n$, ie:

$$\mathfrak{F} =_{df} \{C_S \mid S \in \mathcal{B}([0,1]^n)\}, \text{ where } C_S =_{df} \{G_{\vec{z}} \mid \vec{z} \in S\}$$

Then, clearly, $\mathfrak{F}$ is a σ-algebra, ϕ is a random variable and for every probability measure λ on $\mathcal{B}([0,1]^n)$, $\mu =_{df} \lambda \circ \phi$ is a probability measure on $\mathfrak{F}$. Moreover if there exists some Borel function $h : [0,1]^{n+2} \to \mathbb{R}$ and some Borel set $A \in \mathcal{B}(\mathbb{R})$ such that:

$$(x, y) \in G_{\vec{z}} \Leftrightarrow h(\vec{z}, x, y) \in A \tag{2.70}$$

then (A1), (A2) and (A4) all hold.

Proof. Since $C_S = \phi^{-1}(S)$ by construction, clearly $\mathfrak{F}$ is a σ-algebra and ϕ is a random variable. This same reasoning establishes that $\lambda \circ \phi$ is a probability measure on $\mathfrak{F}$. The truth of the last claim is equally easy to establish, since in the presence of (2.70), D can be written as a composition of h^{-1}, ϕ^{-1} and inverse images of projections, overall acting on some Borel set A. Since all these maps are Borel measurable, so is their composition and hence $D \in \mathfrak{F}$. □

We now provide two examples that satisfy this framework; hence also (A4).

Example 2 (Picking Lines)**.** *Let $S = [0,1]^2$ and let Ω consist of all vertical lines:*

$$\Omega =_{df} \{L_z \mid z \in [0,1]\}, \text{ where } L_z =_{df} \{(z, y) \mid y \in [0,1]\}$$

We then observe that

$$(L_z, (x, y)) \in D \Leftrightarrow z - x = 0$$

Since $h(z, x, y) = z - x$ and $\{0\}$ are both Borel, this example falls under the general framework.

Example 3 (Picking Disks)**.** *Let $S = [0,1]^2$ and let Ω contain disks as follows:*

$$\Omega =_{df} \{B_{r,x',y'} \mid (r, x', y') \in [0,1]^3\},$$

where $B_{r,x',y'} =_{df}$ the disk of radius r centered at (x', y') intersected with S

We now observe that:

$$(B_{r,x',y'}, (x,y)) \in D \Leftrightarrow (x-x')^2 + (y-y')^2 - r \leq 0$$

Since $h(r,x',y',x,y) = (x-x')^2 + (y-y')^2 - r$ *and* $(-\infty, 0]$ *are both Borel, this example falls under the general framework.*

Since (A4) holds of the last two examples, then also (A1) and (A2) hold. What about (A3)? Clearly the answer will depend on the measure on $(\Omega, \mathfrak{F})$, which in turn, via the general framework, depends on the choice of measure λ over the parameter space. In the simplest case where $\lambda =_{df} Leb$, (A3) holds trivially both in the case of lines and that of disks:

$$Pr(x,y) = \mu(\{L_z \mid (x,y) \in L_z\}) = \lambda(\{z \mid x = z)\}) = 0$$

$$\begin{aligned} Pr(x,y) &= \mu(\{B_{r,x',y'} \mid (x,y) \in B_{r,x',y'}\}) \\ &= \lambda(\{(r,x',y') \mid (x-x')^2 + (y-y')^2 - r \leq 0\}) = \pi r^2 \end{aligned}$$

Both these quantities are independent of (x,y), hence trivially Riemann Integrable. This is because our choice of λ was the choice of 'uniform' measure, which is translation invariant. However, the calculations above ought to convince the reader that a wide range of well-behaved choices of measure (for instance, measures induced by continuous distribution functions) will lead to Riemann Integrable expressions for $Pr(x,y)$.

Remark. It is possible to generalise further the framework above, without relapsing to the abstractness of the original problem. For instance, everything works as before if we allow ourselves to use *any* Borel function $h : \mathbb{R}^n \to \mathbb{R}^m$ to capture the relation $(x,y) \in G$, as opposed to restricting attention to Borel functions $h : \mathbb{R}^n \to \mathbb{R}$. This generalisation allows us for instance to deal with *rectangles*:

$$(x,y) \in [a,b] \times [c,d] \Leftrightarrow (a-x, x-b, c-y, y-d) \in (-\infty, 0] \times ... \times (-\infty, 0]$$

where:

$$h(a,b,c,d,x,y) = (a-x, x-b, c-y, y-d)$$

The key then is to pick a parameterisation ϕ that makes it possible for us to produce a toolbox of functions $h : \phi[\Omega] \times S \to \Omega'$ *known to be measurable* with respect to the respective product measure. It falls beyond the scope of this dissertation to discuss such classes of functions other than the Borel functions.

Remark. Observe that we have not produced an example where separate measurability is satisfied but product measurability is not.

2.6 Conditional Probability

We now return to the development of the theory and the question of the correct way to generalise the elementary definition of conditional probability.

Beyond Elementary Conditional Probability - a discussion

The Elementary Definition

The defining equation of elementary conditional probability is the following:

$$P_A(B) =_{df} \text{ the unique solution to '} P_A(B)P(A) = P(B)\text{'} \qquad \text{(Elem)}$$

This fails to give a unique solution for $P_A(B)$ when $P(A) = 0$, since *any* value for $P_A(B)$ would satisfy this equation. By now it should be obvious however that there are plenty of situations where '$P(A) = 0$' is very different from the statement 'A is impossible', which urges us to provide a definition of conditional probability that allows us to condition upon events of probability 0.

The Limit Definition

One's first guess would presumably be that a limit operation may be used to circumvent the 'division by zero' problem, in the spirit of the definition of derivatives in analysis. We would in this approach think of the set A as the limit value of a decreasing sequence of sets $(A_n : n \in \mathbb{N})$ of *non-zero probability*:

$$A = \bigcap_{n\in\mathbb{N}} A_n, \text{ where for all } n,\ P(A_n) > 0$$

and then consider the limit of the values obtained by conditioning in an elementary manner on each set A_n:

$$P_A(B) =_{df} \lim_{n\to\infty} P_{A_n}(B) =_{df} \lim_{n\to\infty} \frac{P(A_n \cap B)}{P(A_n)} \qquad \text{(Lim)}$$

Clearly the problem here is that no canonical choice of a family $(A_n : n \in \mathbb{N})$ suggests itself in general, whereas $P_A(B)$ is not well-defined as it stands since it can be easily seen to *vary* with the choice of $(A_n : n \in \mathbb{N})$. Indeed the Borel paradox of the great circle was seen precisely as a counterexample to this definition by Borel himself, as we will see at the end of this section.

On the other hand, in most questions of geometric probability, the set A we are conditioning on is in fact understood as a member of a partition generated by a certain real parameter, usually a coordinate function. For instance, in the Borel paradox, one wishes to condition on the event that the value of the longitude coordinate function be a. This event, $A =_{df} \{\omega : \phi(\omega) = a\}$, is indeed of measure 0. It is however not understood singly for a particular value of a only, but rather as an element of the set of great circles. As a result, it can be readily seen as the limit of ever thinner 'bundles' of great circles:

$$A = \{\omega : \phi(\omega) = a\} = \bigcap_n \{\omega : \phi(\omega) \in [a, a + 1/n]\}$$

or in equivalent notation:

$$A = \phi^{-1}(\{a\}) = \bigcap_n \phi^{-1}([a, a + 1/n])$$

We can now use $(\phi^{-1}([a, a + 1/n]) : n \in \mathbb{N})$ as the canonical choice in (Lim) to produce a definition of the probability of B conditional on the event $\phi(\omega) = a$:

$$P_{\phi^{-1}(\{a\})}(B) =_{df} \lim_{n\to\infty} \frac{P(\phi^{-1}([a, a + 1/n]) \cap B)}{P(\phi^{-1}([a, a + 1/n]))} \qquad \text{(Rv)}$$

The choice of label for the above equation rests on the immediate observation that this definition can only work when ϕ is a *random variable*, since the inverse image of the intervals $[a, a + 1/n]$ for each n and a must be in $\mathfrak{F}$.

Kolmogorov's Definition

Kolmogorov's approach is somewhat different than the above. It rests however on the same observation that it is impossible to define a notion of probability conditional on *a single event* of probability 0, but it may be possible to define a notion of probability conditional on *the value of a probability function* or, synonymously, on *a choice out of a partition of* Ω *indexed by an arbitrary set* I:

$$A =_{df} u^{-1}(a), \qquad \text{for some } a \in I.$$

Hence, we ought to abandon the type of object employed in the elementary case,

$$A \;\; \mapsto \;\; (\, B \mapsto P_A(B)\,)$$

for the following type of object:

$$u \;\; \mapsto \;\; (\, (B, a) \mapsto P_{u^{-1}(a)}(B)\,)$$

This will be a very different kind of entity than that of elementary conditional probability. We hence replace the notation $P_{u^{-1}(a)}(B)$ with the new notation $P_u(B)(a)$, to keep the two notions separate[42]. The new notation also emphasizes the point that, assuming we can make it precise, the notion denoted by $P_u(B)(a)$ will be a *function* from $u[\Omega]$ to $[0, 1]$, when viewed as a function of a.

Kolmogorov's definition of $P_u(B)$ rests on the requirement that $P_u(B)$ must agree with elementary conditional probability, whenever the latter is defined. Certainly this entails the *coincidence* of the two notions whenever $u^{-1}(a)$ has nonzero probability:

$$\text{if } P(u^{-1}(a)) \neq 0, \;\; \text{then } P_{u^{-1}(a)}(B) = P_u(B)(a)$$

However, our new notion must also agree with the elementary values $P_{u^{-1}[C]}(B)$, whenever they are defined (i.e., whenever $u^{-1}[C]$ has nonzero probability).

[42]Kolmogorov uses $P_u(a; B)$ in place of $P_u(B)(a)$.

This agreement translates to the requirement that the *integral* (expectation) of $P_{u^{-1}(a)}(B)$ over $a \in C$ must yield the same value as the elementary formula:

$$P_{u^{-1}(C)}(B) = \int_C P_u(B)(a)\ d_\mu(a), \quad (P(u^{-1}[C]) \neq 0)$$

where the measure μ we must integrate with respect to is the measure induced over possible values of a, appropriately conditioned (in an elementary manner):

$$\mu =_{df} P^{(u)}_{u^{-1}(C)}$$

Remark. Using formula (2.55) we can drop the notationally awkward dependence of μ on $u^{-1}(C)$ and rewrite the condition of interest as follows:

$$P_{u^{-1}(C)}(B) = \frac{1}{P(u^{-1}(C))} \int_C P_u(B)(a)\ d_{P^{(u)}}(a), \quad \text{where } u^{-1}(C) \neq 0. \quad \text{(Kolm)}$$

It follows from a nontrivial argument which we will produce in the next subsection, that this equation indeed defines $P_u(B)$ to be a *random variable* $u[\Omega] \to [0,1]$, but *only up to equivalence*. In other words, there always exists one random variable $P_u(B)$ that satisfies (Kolm) and any two such random variables are equal everywhere on $a \in u[\Omega]$, except possibly on a certain set of probability 0.

In the context of definition (Rv) above, the set C would be an interval $[a, a + 1/n]$ and the inverse image of C would be a 'bundle' of great circles. In fact, one immediately now suspects that (Rv) may prove to be a special case of (Kolm), since 'integrals' are a stronger existence requirement than 'derivatives'. This is indeed the case, albeit with the unavoidable introduction of the 'almost surely' disclaimer.

This completes our discussion of how one can motivate Kolmogorov's notion from that of elementary conditional probability. It is certainly a natural step to take - although formally it is not a *generalisation* of the elementary notion, but rather a different entity altogether.

Kolmogorov's Definition of Conditional Probability

Definition 40. Any random variable $P_u(B)$ that satisfies (Kolm) is called a *version of the conditional probability of B with respect to the partitioning u*:

$$\forall\, C \in \mathfrak{F}^{(u)}, \quad P(u^{-1}(C) \cap B) = \int_C P_u(B)(a)\; d_{P^{(u)}}(a) \tag{Kolm}$$

Remark. Observe that $P_u(B)$ must be a random variable for its Lebesgue integral to be defined. Observe also that (Kolm) trivially holds if $P(u^{-1}(C)) = 0$, whereas, if not, we can divide both sides by $P(u^{-1}(C))$ to obtain the statement of (Kolm) we encountered in the previous subsection.

We would like to prove two things for Definition 40 to be satisfactory. Firstly, we must establish *existence*: ie that there always exists a random variable that satisfies (Kolm). Indeed, existence is an almost trivial application of the Radon-Nikodým Theorem, as we will see. Secondly, we would *like* to establish *uniqueness*, ie that there only exists *one* random variable that satisfies (Kolm). This is however not true, which is why we have introduced the term 'version of the conditional probability' - the best we can do is establish *uniqueness up to equivalence*, ie that any two versions will be equal for all $a \in u[A]$, except possibly on a set $C \in \mathfrak{F}^{(u)}$ with $P^{(u)}(C) = 0$. We formally state these results in the following theorem:

Theorem 19. *There always exists a random variable $P_u(B)$ that satisfies (Kolm). Moreover, any two such random variables are equivalent.*

Proof. We first establish uniqueness up to equivalence. Consider any two random variables $x : u[\Omega] \to \mathbb{R}$ and $y : u[\Omega] \to \mathbb{R}$ that both satisfy (Kolm) for any $C \in \mathfrak{F}^{(u)}$. Then:

$$\int_C x(a) d_{P^{(u)}}(a) = \int_C y(a) d_{P^{(u)}}(a) = P(B \cap u^{-1}(C))$$

by dividing (Kolm) through by $P^{(u)}(C)$ if it is nonzero, or trivially so otherwise. Therefore:

$$\forall\, C \in \mathfrak{F}^{(u)} : \quad \int_C x(a) d_{P^{(u)}}(a) = \int_C y(a) d_{P^{(u)}}(a)$$

which implies, by (IX) of Theorem 12, that x is equivalent to y, as required.

To establish existence, we need to state the *Radon-Nikodým* theorem ([1]):

Theorem (Radon-Nikodým)**.** *Let μ and λ be σ-finite measures on a σ-algebra Σ associated to a set S, such that*

$$\forall C \in \Sigma, \quad \lambda(C) \neq 0 \;\Rightarrow\; \mu(C) \neq 0; \tag{2.71}$$

then $\lambda = f\mu$ for some non-negative Borel function $f : S \to \mathbb{R}$ (ie a non-negative random variable), where:

$$\lambda = f\mu \text{ means } \lambda(C) =_{df} \int_C f(a) d_\mu(a) \tag{2.72}$$

To use this theorem for our purposes, we prove that the conditions of the theorem hold in the following case:

$$S =_{df} u[\Omega], \quad \Sigma =_{df} \mathfrak{F}^{(u)}$$

$$\mu =_{df} P^{(u)}$$

$$\lambda : C \mapsto P(B \cap u^{-1}[C]), \quad (C \in \mathfrak{F}^{(u)})$$

Certainly μ and λ are σ-finite measures, since probability measures are by definition *finite*, hence also σ-finite. Also λ is a measure: since inverse images commute with all set operations, λ inherits countable additivity from P. Finally, condition (2.71) holds by virtue of the elementary observation:

$$\forall\, C \in \mathfrak{F}^{(u)} : \quad \mu(C) =_{df} P(u^{-1}[C]) \geq P(B \cap u^{-1}[C]) =_{df} \lambda(C)$$

Note that this is a very easy application of the Radon-Nikodým Theorem, since the two measures λ and μ are related to each other in a very simple manner. We have hence established that

$$\forall C \in \mathfrak{F}^{(u)} : \quad P(B \cap u^{-1}(C)) =: \lambda(C) = \int_C f(a) d_\mu(a) \; =_{df} \int_C f(a) d_{P^{(u)}}(a)$$

for some non-negative random variable f, which proves existence. □

We can now investigate whether $P_u(B)(a)$ *as a function of* B satisfies the axioms of probability, as was the case in the elementary definition of $P_A(B)$. We will see that it does 'almost surely' (in the sense of $P^{(u)}$):

Theorem 20. *Almost surely* $0 \leq P_u(B) \leq 1$

Proof. Recall that $P_u(B)$ is almost surely equal to f in the proof of Radon-Nikodým, which is guaranteed to be non-negative. This proves that almost surely $0 \leq P_u(B)$.

We now show that $P_u(B) \leq 1$ almost surely. We assume by way of contradiction that there exists some $M \in \mathfrak{F}^{(u)}$ such that $P^{(u)}(M) > 0$ and $P_u(B)(a) > 1$ for every a in M. We can in fact make this statement stronger. Observe that

$$P_u(B)(a) > 1 \;\Leftrightarrow\; \exists n,\, P_u(B)(a) \geq 1 + 1/n$$

$$\therefore\; M \subseteq \bigcup_n M_n, \text{ where } M_n =_{df} \{a \mid P_u(B)(a) \geq 1 + 1/n\}$$

Hence $P^{(u)}(M_k) > 0$ for at least one natural number k, otherwise

$$P^{(u)}(M) \leq P^{(u)}(\bigcup_{n \in \mathbb{N}} M_n) = \sum_n P^{(u)}(M_n) = 0$$

contradicting our hypothesis that $P^{(u)}(M) > 0$. Hence, letting $M' =_{df} M_k$,

$$\begin{aligned} P(B \cap u^{-1}(M')) &\geq P_{u^{-1}(M')}(B) \text{ by definition of elementary conditional probability} \\ &= E_{u^{-1}(M')}(P_u(B)) \text{ by (Kolm)} \\ &\geq E_{u^{-1}(M')}(1 + 1/n) \text{ by Theorem 13.(II)} \\ &= (1 + 1/n)P(u^{-1}(M')) \text{ by definition of expectations} \\ &> P(u^{-1}(M')), \text{ which is a contradiction.} \end{aligned}$$ □

Theorem 21. *If $B = \bigcup_{n\in\mathbb{N}} B_n$ where the B_n's are pairwise disjoint, then almost surely $P_u(B) = \sum_n P_u(B_n)$.*

Proof. First observe that if we set $C = u[\Omega]$ in (Kolm) we get:

$$P(B) = E(P_u(B)) \tag{2.73}$$

Now we have that:

$$\begin{aligned} P(B) &= \sum_n P(B_n), \text{ by countable additivity of } P \\ &= \sum_n E(P_u(B_n)), \text{ by (2.73)} \\ &= \sum_n E(|P_u(B_n)|), \text{ since } P_u(B_n) = |P_u(B_n)| \text{ a.s., by (VI) of Theorem 13} \end{aligned}$$

So, in particular, this latter expression converges (since $P(B)$ is finite). Then for any $C \in \mathfrak{F}^{(u)}$ such that $P^{(u)}(C) > 0$, we get:

$$\begin{aligned} E_{u^{-1}[C]}(P_u(B)) &= P_{u^{-1}[C]}(B), \text{ by (Kolm)} \\ &= \sum_n E_{u^{-1}[C]}(P_u(B_n)), \text{ by additivity of } P \text{ and (2.73)} \\ &= E_{u^{-1}[C]}\left(\sum_n P_u(B_n)\right), \text{ by (V) of Theorem 13} \end{aligned}$$

since $\sum_n E(|P_u(B_n)|)$ converges. But this latter implies that $\sum_n P_u(B_n) = P_u(B)$ a.s., by the same proof as the one for uniqueness a.s. in Theorem 19. □

The Limit Definition of Conditional Probability

In case the partition is generated by a random variable $x : \Omega \to \mathbb{R}$, a simpler definition is available as we previously discussed:

$$P_x(B)(a) =_{df} \lim_{h\to 0} \frac{P(x^{-1}[a, a + h] \cap B)}{P(x^{-1}[a, a + h])} \tag{Lim}$$

This presupposes that $P(x^{-1}[a, a+h]) \neq 0$ and, of course, that the limit featured exists, two assumptions somewhat hard to work with. We then amend (Lim) as

follows to circumvent them:

$$P_x(B)(a) =_{df} \begin{cases} \text{the limit in (Lim), if it exists} \\ 0, \text{ otherwise} \end{cases} \tag{Lim*}$$

We allow ourselves to use the same notation $P_x(B)$ for the probability function defined by either (Kolm) or (Lim*), precisely because of the following result:

Theorem 22. *Fix a random variable x. Then any random variable that satisfies (Kolm) is almost surely equal to the probability function defined in (Lim*).*

Proof. Let f be a random variable that satisfies (Kolm) and let g be the probability function defined in (Lim*):

$$g : a \mapsto \begin{cases} \text{the limit in (Lim), if it exists} \\ 0, \text{ otherwise} \end{cases}$$

Both f and g are probability functions $x[\Omega] = \mathbb{R} \to \mathbb{R}$. Recall definition (Kolm):

$$\forall\, C \in \mathfrak{F}^{(x)}, \quad P(x^{-1}(C) \cap B) = \int_C f(a)\; d_{P^{(x)}}(a) \tag{Kolm}$$

We now take cases, according to whether $P(B) = 0$. If yes, $g(a) = 0$ for all a, by definition (Lim*). Moreover, (Kolm) becomes:

$$\forall\, C \in \mathfrak{F}^{(x)}, \quad \int_C f(a)\; d_{P^{(x)}}(a) = 0$$

Therefore all the integrals of f with respect to the measure $P^{(x)}$ are equal to the integrals of g (since we have shown g to be identically zero). Hence by IX of Theorem 12, it follows that f is equal to g almost surely, as required.

Now for the nontrivial case, where $P(B) \neq 0$. We rewrite (Kolm) as follows:

$$\forall\, C \in \mathfrak{F}^{(x)}, \quad P(B)P_B(x^{-1}(C)) = \int_C f(a)\; d_{P^{(x)}}(a)$$

using the elementary definition for $P_B(x^{-1}(C))$. Taking $C =_{df} (-\infty, t)$,

$$P(B)F_B^{(x)}(t) = \int_{-\infty}^{t} f(a)\; d_{P^{(x)}}(a)$$

where $F_B^{(x)}$ is the distribution function induced by x from the measure P_B. Kolmogorov now invokes a theorem of Lebesgue's to finally infer that:

$$\text{almost surely } \; f(a) = P(B) \lim_{h \to 0} \frac{F_B^{(x)}(a+h) - F_B^{(x)}(a)}{F^{(x)}(a+h) - F^{(x)}(a)} \tag{2.74}$$

The RHS is of course by definition equal to g iff it exists, so we are done. □

Remark. Definition 40 suffers in intuitive content from the 'almost surely' disclaimer. In questions of geometric probability one certainly expects the theory to provide him with a *number* as the answer to fully specified questions of conditional probability. A real function defined almost everywhere is, in some sense, too abstract an object to replace a mere number as the answer to such a question. However, in light of Theorem 22 one may now choose a simpler route and *define* conditional probability with respect to a random variable *always* via (Lim*), assured that it also possesses the intuitively appealing property (Kolm). In more abstract contexts (notably beyond geometric probability) where x is *not* a random variable, one may use (Kolm) instead. There is perhaps some loss in consistency but considerable gain in intuitive content.

As a final addition to this chapter, we proceed to investigate under which conditions we can rewrite the definition (Lim*) using densities so that no limit operation is involved. We can observe immediately that if the densities $f^{(u)}(a)$ and $f_B^{(u)}(a)$ exist and moreover $f^{(u)}(a) > 0$, then by the definition of probability densities and (Lim*) we evidently get:

$$P_u(B)(a) = P(B)\frac{f_B^{(u)}(a)}{f^{(u)}(a)} \tag{2.75}$$

which we can rewrite as follows:

$$P(B)f_B^{(u)}(a) = P_u(B)(a)f^{(u)}(a) \tag{2.76}$$

It is awkward that we need to posit the existence of both densities as well as that $f^{(u)}(a) > 0$ to obtain this formula. Especially the latter requirement is on the whole quite arbitrary. We conclude this section then by deriving (2.76) under less arbitrary assumptions: the existence of $f^{(u)}(a)$ and of the limit in (Lim):

Proposition 40. *If the limit in (Lim) and the density $f^{(u)}(a)$ both exist, then $f_B^{(u)}(a)$ also exists and satisfies:*

$$P(B)f_B^{(u)}(a) \leq f^{(u)}(a) \tag{2.77}$$

Remark. Note that we cannot derive (2.77) straightforwardly by claiming that $P_u(B)(a) \leq 1$ since we are only assured of this latter fact almost surely.

Proof. Assume the density $f^{(u)}(a)$ exists. Then by definition:

$$f^{(u)}(a) =_{df} lim_{h\to 0}(F^{(u)}(a+h) - F^{(u)}(a))$$

exists. We have also assumed the limit in (Lim) exists, so by standard analysis

the product of the limits is the limit of the products:

$$\Big(lim_{h\to 0}F^{(u)}(a+h) - F^{(u)}(a)\Big) \cdot \left(\lim_{h\to 0} \frac{F_B^{(u)}(a+h) - F_B^{(u)}(a)}{F^{(u)}(a+h) - F^{(u)}(a)}\right) =$$

$$= \lim_{h\to 0} \left((F^{(u)}(a+h) - F^{(u)}(a)) \cdot \frac{F_B^{(u)}(a+h) - F_B^{(u)}(a)}{F^{(u)}(a+h) - F^{(u)}(a)}\right) =$$

$$= \lim_{h\to 0}(F_B^{(u)}(a+h) - F_B^{(u)}(a)) =_{df} f_B^{(u)}(a)$$

Finally (2.77) holds since:

$$P(B)f_B^{(u)}(a) =_{df} P(B)lim_{h\to 0}(F_B^{(u)}(a+h) - F_B^{(u)}(a))$$
$$= lim_{h\to 0}\left(P(B)P_B(u^{-1}[(-\infty, a+h)]) - P(B)P_B(u^{-1}[(-\infty, a)])\right)$$
$$= lim_{h\to 0}\left(P(B\cap u^{-1}[(-\infty, a+h)]) - P(B\cap u^{-1}[(-\infty, a)])\right)$$
$$= lim_{h\to 0}P(B\cap u^{-1}[(a, a+h)]), \text{ since } B\cap u^{-1}[(-\infty, a)] \subset B\cap u^{-1}[(-\infty, a+h)]$$
$$\leq lim_{h\to 0}P(u^{-1}[[a, a+h)]), \text{ since } P(B\cap u^{-1}[[a, a+h)]) < P(\cap u^{-1}[[a, a+h)])$$
$$= f^{(u)}(a) \text{ by the same argument as above} \qquad \square$$

So now we need only observe that in case $f^{(u)}(a) = 0$, by (2.77) also $f_B^{(u)}(a) = 0$, so (2.76) holds in that case, too.

Chapter 3

Resolving the Paradoxes

Finally we can revisit the paradoxes that we discussed in Chapter 1. Trivially we observe that all paradoxes except that of the Great Circle utilise simple calculations of *areas*, based on the assumption that some real parameter, or a pair of them, can be modelled by the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), Leb)$, or respectively the space $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), Leb)$. There is therefore no need to go over these trivial examples, since we would in effect be precisely reproducing the original calculations. As we explained in the Introduction, the paradoxicality was removed *de facto* by Kolmogorov-style probability theory, since there the operation of 'picking at random' is not solely determined by the set we are picking from but also by the probability law that governs this selection - it is, that is, identified with a tuple $(\Omega, \mathfrak{F}, P)$.

Anyway, the paradox of the Great Circle provides instances of all notions of interest which we will translate to the formal language in what follows.

3.1 The Paradox of the Great Circle

Description of the Underlying Probability Space

We model the choice of a point at random from the sphere as follows:

$$\Omega =_{df} \{\text{points on the surface of a unit radius sphere}\} \subseteq \mathbb{R}^3$$
$$\mathfrak{F} =_{df} \{\text{Borel sets on } \Omega\}, \quad P(A) =_{df} Leb(A), \ (A \in \mathfrak{F})$$

where we represent $\mathbb{R}^3$ in the standard *orthocanonical* xyz-coordinate system so that Ω be of unit radius, with its center at the origin, its North Pole set at the point $(0, 1, 0)$ and its South Pole at the point $(0, -1, 0)$.

The Spherical Coordinate System

We now take some time to describe in detail the *spherical* coordinate system, to avoid comfusion in what follows.

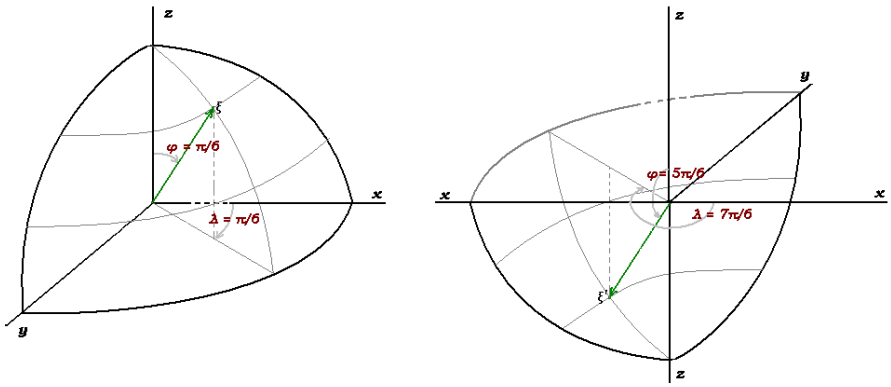

Figure 3.1: A depiction of two points on the sphere and their respective latitudes and longitudes.

Definition 41. We consider the following functions defined on Ω.

$$\phi(\xi) =_{df} \textit{co-latitude of } \xi, \text{ in the range } [0, \pi]$$
$$\lambda(\xi) =_{df} \textit{longitude of } \xi, \text{ in the range } [0, 2\pi)$$

defined as follows:

- Define ϕ to be the (unique acute) polar angle of the position vector of ξ from the positive z-axis.
- Define λ to be the angle in the equatorial plane between the projection of the position vector of ξ and the positive x-axis, measured clockwise if one visualises the equatorial plane from above.

Remark. Observe that the common usage terms 'co-latitude' and 'longitude' do not numerically correspond to the functions we have defined above. For instance, 'longitude' is measured with positive values northwards of the equatorial plane and negative values southwards, which is convenient for geography but not for integration. For this reason, it is standard mathematical practice to abandon them in favor of the conventions above. This is an insignificant matter for our purpose eitherway since circles of latitude and half-meridians are preserved under either convention:

Definition 42. The reader can also refer to Figure 3.2.

- A *meridian*, otherwise known as a *great circle*, is the intersection of Ω with a plane containing the North and South Pole.
- A *half-meridian* is a segment of a meridian that starts from the North Pole and ends at the South. It can be expressed as a set of points of constant longitude $\{\xi \in \Omega \mid \lambda(\xi) = c\}$, where $c \in [-\pi, \pi)$ a half-meridian.

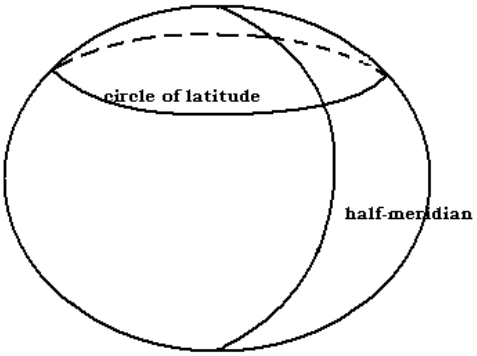

Figure 3.2: A depiction of a circle of latitude and a half-meridian.

- A *circle of latitude* is the intersection of Ω with a plane parallel to the Equator. It can be expressed as a set of points of constant co-latitude $\{\xi \in \Omega \mid \phi(\xi) = c\}$, where $c \in [-\pi/2, \pi/2)$.

Clearly, a choice of half-meridian and a circle of latitude or, equivalently, a choice of co-latitude and longitude, uniquely identify each point on Ω.

The Formal Resolution of the Paradox

Formally ϕ and λ are probability functions carried by $(\Omega, \mathfrak{F}, P)$. A simple calculation establishes that they are in fact random variables.

Proposition 41. *The probability functions ϕ and λ are both random variables.*

Proof. Inverse images of half-rays under either ϕ or λ can be obtained as intersections of simple Borel sets in $\mathbb{R}^3$ by considering the planes that contain the respective circles of latitude or great circles. □

Therefore, as in the section on constructing multidimensional probability spaces via random variables, we induce the probability space $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), P^{(u)})$, where $u(\xi) =_{df} (\phi(\xi), \lambda(\xi))$. The following formula will prove very useful:

Proposition 42. *For any $\phi_1, \phi_2 \in [0, \pi]$, $\lambda_1, \lambda_2 \in [0, 2\pi)$, we have*

$$P^{(u)}([\phi_1, \phi_2] \times [\lambda_1, \lambda_2]) = \frac{1}{4\pi}(\lambda_2 - \lambda_1)(\cos(\phi_1) - \cos(\phi_2)) \tag{3.1}$$

Proof. By the definition of $P^{(u)}$ and P we get that

$$P^{(u)}([\phi_1, \phi_2] \times [\lambda_1, \lambda_2]) = Leb(\{\xi \mid \phi_1 \leq \phi(\xi) \leq \phi_2, \lambda_1 \leq \lambda(\xi) \leq \lambda_2\})$$

But by standard results in real calculus the RHS is equal to:

$$\int_{\phi_1}^{\phi_2} \int_{\lambda_1}^{\lambda_2} \sin\phi \, d\phi \, d\lambda = \frac{1}{4\pi}(\lambda_2 - \lambda_1)(\cos(\phi_1) - \cos(\phi_2)) \qquad \square$$

We now gradually reproduce the argument in the paradox of the great circle:

> Pick two points on the sphere at random. What is the probability that they lie within $10'$ of each other? By symmetry we assume that the first point is fixed on the North Pole of the sphere. We then calculate the proportion of the sphere's surface that lies within $10'$ of the North Pole. This is 2.1×10^{-6}.

We assume at this point that we may indeed fix the first point to be the North Pole (we prove this leads to no loss in generality in the next subsection). Then the angle between a point on the sphere and the North Pole is precisely given by its co-latitude. So we can express formally the event of interest, which we denote by B:

$$B =_{df} \{\xi \mid \phi(\xi) < c\}$$

where in this case, $c = 10' = 2\pi/(6 \cdot 360) = \pi/1080$. The set B is of course the inverse image of $[0, \phi_2] \times [\lambda_1 \times 2\pi]$ under u. We can hence evaluate $P(B)$ via (3.1):

$$P(B) =_{df} P^{(u)}([0, \phi_2] \times [\lambda_1 \times 2\pi]) = \frac{1}{4\pi}(2\pi - 0)(\cos(0) - \cos(c)) = \frac{1 - \cos(c)}{2}$$

which the reader may check evaluates to 2.115397×10^{-6} when $c = 10'$.

We now investigate the alternative calculation that allegedly produces a different value:

> We may however observe that there exists a unique great circle that connects the second randomly selected point with the North Pole. Moreover, by rotational symmetry, no great circle has more chances of being selected than any other. Therefore, we may assume we know the great circle that connects the two points. We have now reduced the original problem to one of picking one point on a given great circle. The answer to the original question can hence be found by calculating the proportion of the length of the great circle that lies within $10'$ of the North Pole, which is of course $2/(360 \cdot 60) \approx 9.3 \times 10^{-4}$, not 2.1×10^{-6}.

Two separate claims are being made in this passage. The first claim is that conditioning on a choice of half-meridian λ_0 leaves the probability of B unaffected. We call this the *symmetry assumption* and formally we write it as follows:

$$P(B) = P_\lambda(\lambda_0; B) \qquad \text{(sym)}$$

where λ is now viewed as a random variable carried by $(\Omega, \mathcal{B}(\Omega), P)$. We now prove that this claim is indeed correct:

Proposition 43. *Assumption (sym) holds.*

Proof. We may assume that $c \neq 0$, since the argument collapses in the (probability 0) eventuality where the two points randomly selected coincide. We then observe that for any $h > 0$ and an arbitrary $\lambda \in [0, 2\pi)$:

$$\begin{aligned} P(\lambda^{-1}([\lambda_0, \lambda_0 + h] \cap B)) &=_{df} P^{(u)}([0, \pi] \times [\lambda_0, \lambda_0 + h] \cap B) \\ &= P^{(u)}([0, c] \times [\lambda_0, \lambda_0 + h]), \text{ by definition of } B \\ &= \frac{1}{4\pi}(\lambda_0 + h - \lambda_0)(\cos(0) - \cos(c)), \text{ by (3.1)} \\ &= \frac{h(1 - \cos(c))}{4\pi} \end{aligned}$$

and similarly

$$P(\lambda^{-1}([\lambda_0, \lambda_0 + h]) = \frac{1}{4\pi}(\lambda_0 + h - \lambda_0)(\cos(0) - \cos(\pi)) = \frac{h}{2\pi}$$

Applying the above to the definition (Lim) of $P_\lambda(\lambda_0; B)$, we obtain that

$$\begin{aligned} P_\lambda(\lambda_0; B) &=_{df} \lim_{h \to 0} \frac{P(\lambda^{-1}([\lambda_0, \lambda_0 + h] \cap B))}{P(\lambda^{-1}([\lambda_0, \lambda + h])} \\ &= \lim_{h \to 0} \frac{1 - cos(c)}{2} = P(B) \end{aligned}$$

as required. □

Remark. Observe that this is in fact a trivial application of (Lim), since no limit operation was actually involved.

The second claim in the quoted passage is that the probability of B conditional on a choice of half-meridian is in fact the 1-dimensional Lebesgue measure of the intersection of B with the half-meridian:

$$P_\lambda(\lambda_0; B) \text{ allegedly is equal to } Leb(B \cap \{\xi \mid \lambda(\xi) = \lambda_0\}) = c/2\pi \qquad \text{(leb)}$$

The reader can check that $c/2\pi$ indeed evaluates to 9.3×10^{-3} when $c = 10'$. However, since we have established (sym) to hold, our formalism automatically rejects (leb), since the value of $P(B)$ is given by the choice of initial probability space and cannot be revised midway. This is what we called a *de facto banishment of the paradox* in the Introduction.

The Resolution of the Paradox - a discussion

The fundamental error in the informal syllogism is that it performs a revision of the probability space midway. A moment's thought can reveal the intuitions that back this revision up: an ill-advised application of the Principle of Indifference together with a misunderstanding of the notion of conditioning on probability 0 events.

On the one hand, the correct intuition captured in (sym) precisely assures us that $P(B)$ ought be equal to $P_\lambda(\lambda_0; B)$. Well, since we know $P(B)$ this immediately determines also $P_\lambda(\lambda_0; B)$, which yields an observation that might be surprising to someone accustomed to the principle of indifference, but is certainly not a paradox:

> The probability of B conditional on a certain fixed choice of half-meridian is in fact not the Lebesgue measure of the arc.

We now need to *separately* assume, on independent grounds, that the conditional probability is in fact given by the length of the arc, so as to turn this into a paradox, as Kolmogorov himself clarifies in [6, p.51]:

> If we assume that the probability distribution of ϕ 'with the hypothesis that ξ lies on the given meridian circle' must be uniform, then we have arrived at a contradiction.

Indeed, to a 19th century probabilist there were convincing independent grounds in favour of (leb). One was inclined to extrapolate from the discrete case that

$$\begin{aligned} Pr(\xi \in B \text{ given } \lambda(\xi) = \lambda_0) &= \frac{Pr(B \cap \{\xi \mid \lambda(\xi) = \lambda_0\})}{Pr(\{\xi \mid \lambda(\xi) = \lambda_0\}} \\ &= \frac{Pr(\text{arc})}{Pr(\text{half-meridian})}. \end{aligned}$$

Now Pr corresponds of course to Lebesgue measure. A naïve substitution then indeed yields:

$$\frac{Leb(\text{arc})}{Leb(\text{half-meridian})} = c/2\pi \qquad \text{(leb)}$$

But 'Pr' corresponds to *2-dimensional* Lebesgue measure, not 1-dimensional Lebesgue measure. Therefore, the notion elementary conditional probability employed above fails, since $Leb(\text{half-meridian}) = 0$, forcing a division by zero.

In the above sense, the temptation to employ the Lebesgue measure as the probability of the arc is not merely an instance of the Principle of Indifference at work, but also a failing of our intuition to accept that lengths are not additively related to areas. It hence becomes important, in geometric probability, to drive the point home that elementary conditional probability is not applicable when conditioning with respect to events of probability 0 (such as a set of lower dimension).

The Choice of Limit

Strictly speaking, the inapplicability of elementary conditional probability *suffices* to resolve the paradox; a novel notion of conditioning with respect to probability 0 events is not required[1]. Both Borel and Kolmogorov understand this well. Borel exclaims ([11, p.19]):

[1] Although in the absence of such a notion we are equally unable to discuss the truth of our correct guess, assumption (sym).

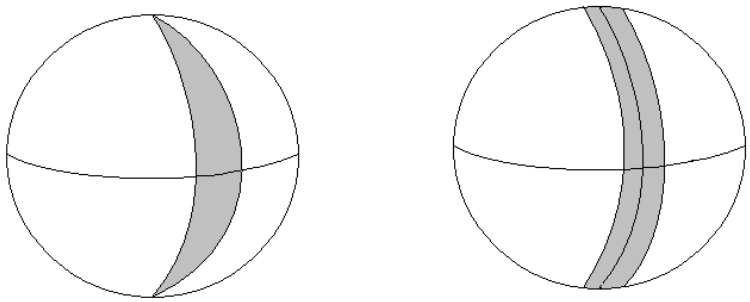

Figure 3.3: On the left, Borel's bundles and the sphere as a union of meridians. On the right, the sphere as a union of circles parallel to a given meridian.

> If the arcs have no width, then in order to speak rigorously, we must assign the value zero to the probability that M and M' are on the circle.

Kolmogorov similarly points out ([6, p.50]) that

> [...]the concept of a conditional probability [in the elementary sense] with regard to an isolated given hypothesis whose probability equals 0 is inadmissible.

However, they both take the opportunity to suggest an alternative way to deal with the problem. Borel suggests precisely what we have been referring to as the *limit definition* of conditional probability ([11, p.19]):

> In order to avoid this factor of zero, which makes any calculation impossible, one must consider a thin bundle of great circles all going through M [the North Pole], and then it is obvious that there is a greater probability for M' [the randomly chosen point] to be situated in a vicinity 90 degrees from M than in the vicinity of M itself.

Borel's understanding of conditional probability as a limit is, as we explained in the introduction to Section 2.6, not admissible without further specifications, since different choices of limit operations could lead to different answers. For instance, different results would be reached were the given meridian the limit of an ever thinner bundle of great meridians or the limit of an ever thinner bundle of circles parallel to the given meridian, otherwise known as a *zone*. The two different limit operations are depicted in Figure 3.1 and the fact that the latter produces a different result can be established by the following simple calculation, resting on the standard fact that the area of spherical zone in a unit sphere is $2\pi h$:

$$\lim_{h\to 0}\frac{P(\text{zone of height } h\cap B)}{P(\text{zone of height } h)}\approx\lim_{h\to 0}\frac{c/2\pi\cdot 2\pi h}{2\pi h}$$
$$=\lim_{h\to 0}\frac{c}{2\pi}=\frac{c}{2\pi}$$

Remark. Observe that we have obtained in this manner (leb). This explains how the paradox of the Great Circle can be seen as a counterexample of the validity of the naïve definition of conditional probability as a limit.

In my opinion, however, the limit of ever thinner zones does not fit the description of the problem, since at any given instance, a zone cannot be expressed in terms of the choice of longitude. In a sense, the way Borel puts it makes it perfectly clear that the choice of limit operation is dictated by the problem itself, by exactly the same token as Kolmogorov's choice of partition is dictated by the problem ([6, p.51]):

> For we can obtain a probability distribution for [the co-latitude ϕ] [...] on the meridian circle only if we regard this circle as an element of the decomposition of the entire spherical surface into meridian circles with the given poles.

Addendum: can we choose where the North Pole is?

Before we proceed to the conclusion of the dissertation, we briefly establish in this subsection that we may indeed fix one point to be the North Pole without loss of generality. Since the two choices are being made independently, the probability space that models the random selection of two points is given by $(\Omega', \mathfrak{F}', P')$, where

$$\Omega' =_{df} \Omega\times\Omega,\ \mathfrak{F}' =_{df} \mathcal{B}(\Omega)\times\mathcal{B}(\Omega),\ P' =_{df} P\times P)$$

We now need to formally express the event "the two points selected lie within c of each other". We define the following function:

$$f(\xi_1,\xi_2) = \text{ angle of shortest arc from } \xi_1 \text{ to } \xi_2$$

Clearly f is well-defined and $f(\xi_1,\xi_2)\in[0,\pi]$. We now show that it is a random variable when seen as a probability function on $(\Omega', \mathfrak{F}', P')$.

Proposition 44. *The probability function f is a random variable.*

Proof. Observe that $(\Omega', \mathfrak{F}', P')$ can be naturally endowed in $(\mathbb{R}^6, \mathcal{B}(\mathbb{R}^6), Leb_6)$. All we need to establish then is that $f(\xi_1,\xi_2)<c$ is a Borel relation in $\mathbb{R}^6$.

All wee need to do that is elementary results from euclidean geometry. Denote the euclidean distance between points ξ_1 and ξ_2 on Ω by $d(\xi_1,\xi_2)$. Then

$$d^2(\xi_1,\xi_2) = 1^2+1^2-2\cdot 1\cdot 1\cdot\cos(f(\xi_1,\xi_2)),\ \text{by the cosine rule}$$

$$\therefore\quad \cos(f(\xi_1,\xi_2)) = \frac{2-d^2(\xi_1,\xi_2)}{2} \tag{3.2}$$

We now observe that

$$\begin{aligned} f(\xi_1,\xi_2) < c &\Leftrightarrow \cos(f(\xi_1,\xi_2)) > \cos(c), \text{ since cos is decreasing in } [0,\pi] \\ &\Leftrightarrow 2 - d^2(\xi_1,\xi_2) > 2\cos(c), \text{ by (3.2)} \\ &\Leftrightarrow 2 - \sqrt{(a_1-a_2)^2+(b_1-b_2)^2+(c_1-c_2)^2} > 2\cos(c) \end{aligned}$$

which certainly defines a Borel relation on the coordinates (a_i, b_i, c_i) of ξ_i. □

Finally, we can establish that we are in fact allowed to fix the first point to be the North Pole, which can be formally expressed as follows.

Proposition 45. *The following holds for all $c \in [0,\pi)$:*

$$P'(f^{-1}(-\infty,c)) = P(\phi^{-1}(-\infty,c)) = \frac{1-\cos c}{2}$$

Proof. Recall that in the previous section we showed $P(\phi^{-1}(-\infty,c)) = \frac{1-\cos c}{2}$ by (3.1). We will show the same of $P'(f^{-1}(-\infty,c))$.

Now consider $\chi(\xi_1,\xi_2)$, the indicator function of the relation "$f(\xi_1,\xi_2) < c$". By the previous proposition, χ is measurable in the product sense, so, by the basic properties of the Lebesgue integral we get:

$$P'(f^{-1}(-\infty,c)) = \int_C \chi(\xi_1,\xi_2) d_{P'}(\xi_1,\xi_2), \quad \text{where } C =_{df} (-\infty,c).$$

We now apply Fubini's Theorem to obtain:

$$P'(f^{-1}(-\infty,c)) = \int_\Omega I_1^\chi(\xi_1) d_P(\xi_1) \tag{3.3}$$

where $I_1^\chi(\xi_1)$ is given by:

$$I_1^\chi(\xi_1) =_{df} \int_\Omega \chi(\xi_1,\xi_2) d_P(\xi_2) = P(\{\xi_2 \mid f(\xi_1,\xi_2) < c\})$$

The RHS is precisely the Lebesgue measure of a spherical cup of angle c, which of course is equal to $P(\phi^{-1}(-\infty,c)) = (1-\cos c)/2$, since the inverse image of a half-ray under ϕ is precisely a spherical cup of angle c. So (3.3) becomes:

$$\begin{aligned} P'(f^{-1}(-\infty,c)) &= \int_\Omega I_1^\chi(\xi_1) d_P(\xi_1) \\ &= \int_\Omega \frac{1-\cos c}{2} d_P(\xi_1) \\ &= \frac{1-\cos c}{2} \int_\Omega d_P(\xi_1) = \frac{1-\cos c}{2}, \text{ as required.} \end{aligned}$$

□

3.2 Conclusion

Kolmogorov's *Foundations of the Theory of Probability* laid the basis for an incredibly powerful axiomatic framework for probability theory. Its significance lies in:

A. Its choice of the triple $(\Omega, \mathfrak{F}, P)$, a measure-theoretic construct, as the fundamental notion.

B. Its technical richness, which allows for advanced techniques such as infinite-dimensional probability spaces (that eventually led to the theory of stochastic processes) and conditional probability with respect to probability 0 events.

We have overviewed in full detail all aspects of the Grundbegriffe except its treatment of infinite-dimensional probability spaces and we have placed its results in a modern measure-theoretic context, amending, completing or replacing proofs as necessary. We have found that in most occasions the modern approach is more streamlined, but in several cases Kolmogorov's approach retains more of the fundamental intuitions underpinning the measure-theoretic recasting of probability.

Such an instance was Kolmogorov's somewhat forlorn treatment of (Riemann) integration of expectations with respect to a parameter, which we have extensively analysed and compared to the closest popular modern result, the Fubini Theorem, in the context of the intriguing experiment of randomly choosing regions from the plane.

Finally, we motivated, developped and contrasted Kolmogorov's notion of conditional probability with respect to probability 0 events with Borel's hints towards a definition of that notion as a limit. We applied these ideas to explain the tangled intuitions underlying the paradox of the Great Circle:

- a confusion between 2-dimensional and 1-dimensional Lebesgue measure,
- a mistaken application of elementary conditional probability,
- an ill-advised application of the Principle of Indifference.

Afterword

As a foreword to this conclusion, we must first and foremost stress that there are several other axiomatisations of probability theory, some of which are very popular. In particular, the Bayesian 'degree of belief' paradigm has been axiomatised by De Finetti as well as Cox and has a great following, in particular among statisticians and Artificial Intelligence scientists.

Despite this plurality, the measure-theoretic paradigm is still considered by most to accurately capture the mathematical content of probability theory, as points out one of the main proponents of the Bayesian paradigm, Edward T. Jaynes in his seminal work, *Probability: the logic of science* ([5]):

> Our system of [Bayesian] probability could hardly be more different from that of Kolmogorov, in style, philosophy and purpose. Yet when all is said and done we find ourselves, to our own surprise, in agreement with Kolmogorov and in disagreement with his critics, on nearly all technical issues.

This story of nearly unequivocal success is an essential component to the resolution of Bertrand paradoxes by Kolmogorov's foundations: on the one hand, feature [A] ensures that protection against Bertrand-type paradoxes is *built in* the formalism. On the other hand, its mathematical success alluded to in feature (B) overwhelms the reader into believing that the paradoxes must be indeed ill-phrased, if they cannot be expressed in such a potent system.

In this sense, Kolmogorov's formalism performs a *de facto* resolution of the paradoxes, without providing any criterion whatsoever for the reader to *choose* a certain one among the different calculations/probability spaces as the most *natural choice* for the particular problem.

The Method of Transformation Groups

Other authors have since attempted to provide such semi-formal criteria. The most notable of these attempts is the Method of Transformation Groups, which has been mostly argued for by Edward T. Jaynes. The general method is to define a group of transformations under which the as of yet unknown probability measure must remain invariant.

This method has been extensively studied independently of the context of Bertrand paradoxes and is the cornerstone of the modern study of geometric probability, known as *integral geometry*[2]. Jaynes however ([5]) proposed to apply it to Bertrand paradoxes by insisting that *any quantity with respect to which the statement of the problem is indifferent, must be an invariant under the choice of probability measure*. In a sense, Jaynes proposed a revision of the Principle of Indifference, whereby indifference no longer corresponds to the choice of uniform measure, but rather to the enforcement of an invariance constraint.

In his pointedly titled paper *The Well-Posed Problem* ([4]), he applies this principle to the Paradox of the Chord with success, uniquely identifying the uniform distribution over the distance between the midpoint of the chord and the center of the disk as the correct choice of measure[3], which he then proceeds to verify experimentally. On the other hand, in other occasions the method fails to work as smoothly, producing "a whole range of choices in some problems, and no [choice of measure][...] free from all objections in others" ([2]).

The method of transformation groups has been the subject of intense study, by Bayesian probability theorists and geometers alike. It becomes of particular relevance in the context of identifying the right choice of *prior distributions* in problems of inference, which is one of the most strongly debated topics in statistics and machine learning.

[2] It is also related to *Felix Klein*'s *Erlangen Program* - see [13].

[3] This calculation appears second in our exposition of the paradox in the Introduction.

Appendix A

Note on Projections and Product Spaces

In constructing a product space out of two spaces $(\Omega_1, \mathfrak{F}_1, P_1)$, $(\Omega_2, \mathfrak{F}_2, P_2)$, we obtain a probability space $(\Omega_1 \times \Omega_2, \mathfrak{F}, P)$ such that the following holds:

$$\mathfrak{F}_i \subseteq \mathfrak{F}^{(\pi_i)} \text{ and } P^{(\pi_i)} = P_i \text{ on } \mathfrak{F}_i$$

Hereby agreement is interpreted as *containment*, rather than *identity*. This is the best we can do in general. Nevertheless, in some cases, notably when $\mathfrak{F}_1 = \mathfrak{F}_2 = \mathbb{R}$, identity is attainable. We proceed to establish this fact, which rests on the topological properties of the reals:

Theorem 23. *Let* $\mathfrak{F} =_{df} \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$. *Then*

$$\mathfrak{F}^{(\pi_1)} = \mathfrak{F}^{(\pi_2)} = \mathcal{B}(\mathbb{R})$$

Proof. We show that $\mathfrak{F}^{(\pi_1)} = \mathcal{B}(\mathbb{R})$. The case $i = 2$ follows by symmetry. Since we know that $\mathfrak{F}_1 \subseteq \mathfrak{F}^{(\pi_1)}$ (see main text), we need only show the converse:

$$A \times \mathbb{R} \in \mathcal{B}(\mathbb{R}^2) \;\Rightarrow\; A \in \mathcal{B}(\mathbb{R}) \tag{A.1}$$

Now fix a constant $c \in \mathbb{R}$ and consider the function $f : \mathbb{R} \to \mathbb{R}^2$, given by

$$f(x) =_{df} (x, c)$$

We readily observe that $f^{-1}(A \times \mathbb{R}) = A$. Therefore for (A.1) it suffices to show that f is Borel measurable. Equivalently, using a Lemma referred to often in the main text, it is sufficient to establish that f is continuous. Let $A \in \mathcal{T}(\mathbb{R}^2)$, an arbitrary open set of $\mathbb{R}^2$. Then we know we can write it as a countable union of open rectangles, as follows:

$$A = \bigcup_{i \in \mathbb{N}} A_i \times B_i$$

$$\therefore\; f^{-1}(A) = \bigcup_{i \in \mathbb{N}} f^{-1}(A_i \times B_i)$$

since the inverse image operator commutes with countable unions (Section 2.1). But we now observe that $f^{-1}(A_i \times B_i)$ is always an open set:

$$\{x \mid x \in f^{-1}(A_i \times B_i)\} = \{x \mid f(x) \in A_i \times B_i\}$$
$$= \{x \mid x \in A_i,\ c \in B_i\} = \begin{cases} \emptyset, \text{ if } c \notin B_i \\ A_i, \text{ otherwise} \end{cases}$$

where both $\emptyset$ and the interval A_i are open sets. □

We are now in a position to prove a statement made in Section 2.1, when we first introduced the construction $\mathfrak{F}^{(u)}$; namely, that it is not always true that $u[A] \in \mathfrak{F}^{(u)}$ for all $A \in \mathfrak{F}$. The standard counterexample to this implication can be obtained for $u =_{df} \pi_1 : \mathbb{R}^2 \to \mathbb{R}$, albeit in a nontrivial manner.

Proposition 46. $A \in \mathfrak{F} \not\Rightarrow u\,[A] \in \mathfrak{F}^{(u)}$

Proof. Fix the space $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), Leb)$ and consider the projection map π_1 defined via $\pi_1(x, y) =_{df} x$. It is then a standard result that

Lemma (Suslin)**.** *Not all projections of Borel sets are Borel.*

Proof of Lemma. Omitted, can be found in [12]. □

Now note that by Theorem 23, $\mathfrak{F}^{(u)} = \mathcal{B}(\mathbb{R})$. Therefore, any Borel set A whose projection is not Borel is precisely an instance where $A \in \mathfrak{F}$ but $\pi_1[A] \notin \mathfrak{F}^{(u)}$, as required for a counterexample. The existence of such an A is precisely guaranteed by the Lemma. □

Remark. The Lemma alluded to above has an interesting history, since Lebesgue himself offered a mistaken proof of its negation in his seminal 1905 monograph [7]. The mistake was spotted in 1916 by *Michail Suslin*, who was then only 19 years old and under the supervision of *Nicolai Lusin*. Suslin proceeded to prove the negation of Lebesgue's statement, offering the proof referred to above.

Note that the erroneous argument had been offered 'on the fly', so to speak, by Lebesgue and did not compromise any of his 1905 results.

Appendix B

Null Sets and Completions

One of the most instrumental notion in applications of lebesgue measure (and measure theory in general) is that of a set of measure 0, otherwise known as a *null* set. This notion allows the formulation of 'almost everywhere' results, whereby a property can be seen to hold for all $x \in \mathbb{R}$ *except* on a null set. In a very real sense, null sets of exceptions may be ignored, which makes the correct formalisation of what precisely a null set is an indispensable tool. Crucially, any such formalisation of null sets is expected to satisfy the following property:

$$A \text{ is null and } C \subseteq A \text{ implies that } C \text{ is null} \tag{B.1}$$

This property is important because in the semantics of null sets as 'sets of exceptions', it would be arbitrary to allow ourselves to exempt a certain set A, but disallow it for a strict subset of A. However, it is clear that in the formulation of measure so far, it may very well be that there exists a certain $A \in \mathfrak{F}$ with $P(A) = 0$. This produces the need for the following definition:

Definition 43. Consider a measure space $(\Omega, \mathfrak{F}, \mu)$. A set A is *null* iff $A \in \mathfrak{F}$ and $\mu(A) = 0$. The space is *complete* iff it satisfies (B.1).

It is an important fact that $((0,1], \mathbf{B}((0,1]), Leb)$ is not a complete space:

Proposition 47. *$((0,1], \mathbf{B}((0,1]), Leb)$ is not complete.*

Proof. □

It turns out however that we can always extend a measure space to a complete space, as follows:

Proposition 48 (Completion of Measure)**.** *Let $(\Omega, \mathfrak{F}, \mu)$ be a measure space. Then consider the family:*

$$\mathfrak{F}^* =_{df} \{F \subseteq \Omega \mid \exists E, G \in \mathfrak{F} \text{ such that } E \subseteq F \subseteq G \text{ and } \mu(G \setminus E) = 0\}$$

Then $\mathfrak{F}^*$ is a σ-algebra and there exists a unique measure μ^* on $\mathfrak{F}^*$ that extends μ on $\mathfrak{F}$ and satisfies (B.1). We call this measure the completion *of μ.*

Proof. It is straightforward to establish that:

$$\mathfrak{F}^* = \sigma(\mathfrak{F} \cup \text{ null sets })$$

We then extend μ to a measure μ^* on $\mathfrak{F}^*$ by setting:

$$\forall F \in \mathfrak{F}^*,\ \mu(F) =_{df} \mu(G) = \mu(E)$$

where we observe that the choice for each F of the 'witnesses' $E, G \in \mathfrak{F}$ is inessential, since any two pairs E_1, G_1, E_2, G_2 that satisfy:

$$E_i \subseteq F \subseteq G_i \text{ and } \mu(G_i \setminus E_i) = 0$$

can be easily shown to have equal measure $\mu^* =_{df} \mu(E_1) = \mu(G_1) = \mu(E_2) = \mu(G_2)$. We can now check without problems that μ^* is indeed a measure on $\mathfrak{F}^*$. Finally, it is unique since it must agree with μ everywhere on $\mathfrak{F}$ and take the value 0 on all null sets; therefore, its values are constrained everywhere on the set $\mathfrak{F} \cup \{$ null sets $\}$, which is a π-system (since the intersection of an arbitrary set with a null set yields a null set). □

In measure theory it is standard to reserve the term *Lebesgue measure* for the *completion* of the space $((0,1], \mathbf{B}((0,1]), Leb)$, whose existence we established in the main text. Naturally most results that hold for the incomplete space also hold for the complete space 'almost everywhere'. As mentioned in the main text, however, probability theorists prefer to work with (incomplete) Borel σ-algebras of topological spaces unless it is necessary for the mathematics to move to their completions. This only becomes necessary in more advanced probability theory than we will encounter in this dissertation.

Remark. It is perhaps surprising that probability theorists do not understand the semantics of 'probability 0 events' to force (B.1) inasmuch as the semantics of 'null sets' do. Certainly the frequentist interpretation seems to suggest (B.1); it seems however that this question has been pushed aside for now, there being too many fundamental difficulties in the interpretation of probability theory that have to be resolved first.

Appendix C

The Shorokhod Representation Theorem

Here, we establish that the conditions are (2.29)-(2.32) are not only necessary for a function to be the distribution of some random variable, but they are also sufficient:

Theorem 24 (Shorokhod Representation)**.** *Let F be any function that satisfies (2.29)-(2.32). Then the function $X_F : [0,1] \to \mathbb{R}$ defined by:*

$$X_F(\omega) := \inf\{z : F(z) > \omega\}$$

is a random variable with distribution function F when understood as a probability function from the space $([0,1], \mathcal{B}([0,1]), Leb)$. We call X_F the Skorokhod Representation[1] *of F.*

Proof. We first establish that X_F is a random variable. Observe that since F is monotonically increasing we have that:

$$\{z : F(z) > \omega\} \text{ is upwards closed}$$

$$\{z : F(z) > \omega\}^c = \{z : F(z) \leq \omega\} \text{ is downwards closed}$$

So:

$$a \in \{z : F(z) > \omega\},\ b \in \{z : F(z) \leq \omega\} \Rightarrow\ a < b$$

and therefore:

$$X_F(\omega) := \inf\{z : F(z) > \omega\} = \sup\{z : F(z) \leq \omega\}$$

[1]Had we assumed that distributions are right-continuous, X_F would have to defined as $\inf\{z : F(z) \geq \omega\}$ for an analogous proof to go through. In any case, the two versions of X_F can be easily seen to agree almost everywhere.

Now for any $a \in \mathbb{R}$:

$$\begin{aligned} a < X_F(\omega) &\Rightarrow a < \inf\{z : F(z) > \omega\}, \text{ by definition of } X_F \\ &\Rightarrow a \notin \{z : F(z) > \omega\}, \text{ by definition of sup} \\ &\Rightarrow F(a) \leq \omega \end{aligned} \tag{C.1}$$

Now let $(a_n : n \in \mathbb{N})$ be an increasing sequence that tends to $X_F(\omega)$. Then:

$$\begin{aligned} &\forall n,\ F(a_n) \leq \omega, \text{ by (C.1)} \\ \therefore\ &\lim_n F(a_n) \leq \omega, \text{ by elementary real analysis} \\ \therefore\ &F(X_F(\omega)) \leq \omega, \text{ since } X_F(\omega) := \lim_n a_n \text{ and } F \text{ is left-continuous} \end{aligned} \tag{C.2}$$

So:

$$\begin{aligned} X_F(\omega) \geq a &\Rightarrow F(X_F(\omega)) \geq F(a), \text{ since } F \text{ monotonic increasing} \\ &\Rightarrow \omega \geq F(a), \text{ by (C.2)} \end{aligned}$$

Conversely:

$$\begin{aligned} \omega \geq F(a) &\Rightarrow a \in \{z : F(z) \leq \omega\} \\ &\Rightarrow a \leq \sup\{z : F(z) \leq \omega\} =_{df} X_F(\omega) \end{aligned}$$

Therefore:

$$\begin{aligned} &X_F(\omega) \geq a \Leftrightarrow \omega \geq F(a) \\ \therefore\ &X_F(\omega) < a \Leftrightarrow \omega < F(a) \\ \therefore\ &X_F^{-1}((-\infty, a)) = \{\omega : \omega < F(a)\} =_{df} (-\infty, F(a)) \cap [0,1] \in \mathcal{B}[0,1] \end{aligned}$$

Recall that the value of the distribution function of X_F at a is defined to be the probability of the inverse image of the half-ray $(-\infty, a)$ under X_F. But we have just shown this inverse image to be $[0, F(a))$, whose probability in the space we are working in is $Leb([0, F(a))) = F(a)$. As a result, F must be equal everywhere to the distribution function of X_F, as required. □

Bibliography

[1] H. Bauer. *Probability Theory and Elements of Measure Theory.* Academic Press, 1981.

[2] A.P. Dawid. Invariant prior distributions. *Encyclopaedia of Statistical Sciences*, 4:228–236, 1983.

[3] Taylor Emmett. *No Royal Road: Luca Pacioli and his Times.* Chapel Hill: the University of North Carolina Press, 1942.

[4] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, sec-4, no 3:227–241, 1968.

[5] E.T. Jaynes. *Probability: the logic of science.* Cambridge University Press, 2003.

[6] A.N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publishing Company, New York, 1956. Translated by Morrison,N.

[7] H. Lebesgue. Sur les fonctions represéntables analytiquement. *Journal de Mathématiques 6ème série*, 1:139–216, 1905.

[8] W. Rudin. *Principles of Mathematical Analysis.* McGraw Hill International Book Company, 3rd Edition, 1976.

[9] W. Rudin. *Real and Complex Analysis.* McGraw Hill Book Company, 3rd Edition, 1987.

[10] G. Shafer. The early development of mathematical probability. Essay available online, 2005.

[11] G. Shafer and V. Vovk. The origins and legacy of kolmogorov's *Grundbegriffe.* Working Paper 4, 2005.

[12] M. Suslin. Sur une definition des ensembles measurables b sans nombres transfinis. *Comptes Rendus Acad. Sci. Paris*, 164:88–91, 1917.

[13] Bas C. Van Fraassen. *Laws and Symmetry.* Oxford University Press, 1989.

[14] J. Von Plato. *Creating Modern Probability.* Cambridge University Press, 1994.

[15] D. Williams. *Probability With Martingales*. Cambridge University Press, 1991.