1

Semantics of Probabilistic Programming: A Gentle Introduction

Fredrik Dahlqvist and Alexandra Silva
University College London
Dexter Kozen
Cornell University

Abstract: Reasoning about probabilistic programs is hard because it compounds the difficulty of classic program analysis with sometimes subtle questions of probability theory. Having precise mathematical models, or *semantics*, describing their behaviour is therefore particularly important. In this chapter, we review two probabilistic semantics. First an operational semantics which models the local, step-by-step, behaviour of programs, then a denotational semantics describing global behaviour as an operator transforming probability distributions over memory states.

1.1 Introduction

A *probabilistic program* is any program whose execution is probabilistic. This usually means that there is a source of randomness that allows weighted choices to be made during execution. Given an initial machine-state, in the event that the program halts, there will be a distribution describing the probability of output events. Any deterministic program is trivially a probabilistic program that does not make any random choices. The source of randomness is typically a *random number generator*, which is assumed to provide independent samples from a known distribution. In practice, these are often *pseudo-random number generators*, which do not provide true randomness, but only an approximation; however, it is possible to construct hardware random number generators that provide true randomness, for example by measuring a noisy electromagnetic process.

Reasoning about deterministic programs usually involves answering binary yes/no questions: *Is the postcondition always satisfied? Does this program halt on all inputs? Does it always halt in polynomial time?* On the other hand, reasoning about probabilistic programming usually involves more *quantitative* questions: *What is the probability that the postcondition is satisfied? What is the probability that this*

[a] From *Foundations of Probabilistic Programming*, edited by Gilles Barthe, Joost-Pieter Katoen and Alexandra Silva published 2020 by Cambridge University Press.

program halts? Is its expected halting time polynomial? In order to answer questions like these, the first step should be to develop a formal mathematical semantics for probabilistic programs, which will allow us to formalise such questions precisely. This is the main purpose of this chapter.

Reasoning about probabilistic programs is in general difficult because it compounds the difficulty of deterministic program analysis with questions of probability theory, which can sometimes be counterintuitive. We will use examples to illustrate all the main ideas presented in this chapter. We introduce these examples here and will return to them as we develop the semantics of probabilistic programs. We start with two examples involving *discrete probabilities* for which naive probability theory provides a sufficient framework for reasoning. We will then present two programs that involve *continuous* distributions for which a more general theory known as *measure theory* is needed. The requisite background for understanding these concepts is presented in Section 1.2.

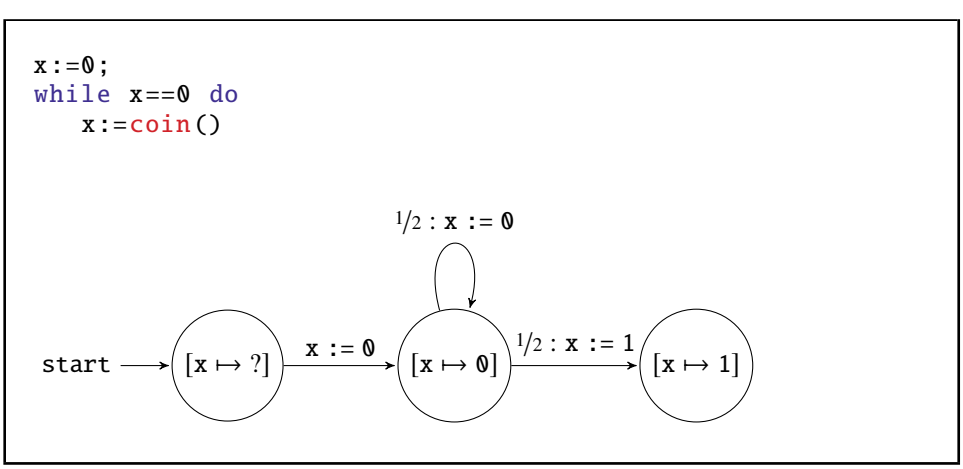

Figure 1.1 A simple coin-toss program

We start with the simple program of Fig. 1.1 displayed next to the small probabilistic automaton it implements. Here the construct `coin()` is our random number generator; each successive call returns 0 or 1, each with probability $1/2$, and successive calls are *independent*, which means that n successive calls will yield one of the 2^n possible sequences of n binary digits, each with probability 2^{-n}. A distribution on $\{0, 1\}$ that takes value 1 with probability p and 0 with probability $1 - p$ is called a *Bernoulli distribution with (success) parameter* p. Thus `coin()` is a Bernoulli distribution with success parameter $1/2$.

It is intuitively clear that this program eventually halts with probability 1. Looking at the automaton of Fig. 1.1, one can see that the probability of the program going

through n iterations of the body of the loop is 2^{-n}. Moreover, the expected number of iterations of the body of the loop is given by

$$\sum_{n=1}^{\infty} n2^{-n} = 2.$$

This type of simple probabilistic process involving repeated independent trials until some fixed "success" event occurs is called a *Bernoulli process*. If the probability of success in each trial is p, then the expected time until success is $1/p$. In this example, $p = 1/2$. We will show in Section 1.3 how the mathematical interpretation of this program (its *semantics*) can be constructed *compositionally*, that is to say line-by-line, and how it agrees with these simple observations.

Our second example is also discrete, but intuitively less obvious. The program of Fig. 1.2 implements a random walk on the two-dimensional grid $\mathbb{Z} \times \mathbb{Z}$. In each iteration of the body of the loop, the function `step` updates the current coordinates by moving left, right, down, or up, each with equal probability $1/4$.

```
main{
    u:=0;
    v:=0;
    step(u,v);
    while u!=0 || v!=0 do
        step(u,v)
}

step(u,v){
    x:=coin();
    y:=coin();
    u:=u+(x-y);
    v:=v+(x+y-1)
}
```

Figure 1.2 A random walk on a two-dimensional grid

The loop continues until the random walk returns to the origin. The first call to `step` outside the loop ensures that the program takes at least one step, so it does not halt immediately. The question of the halting probability is now much less obvious. The state space is infinite, and there is no constraint on how far the random walk can travel from the origin. Indeed, for any distance, there is a nonzero probability that it goes at least that far. However, it turns out that the probability that the program halts is 1. In the terminology of probability theory, we would say that the two-dimensional random walk is *recurrent* at every point. This example illustrates how the analysis of probabilistic programs can rely on results from probability theory

that are far from obvious. Indeed, the three-dimensional version is not recurrent; the probability that a random walk on $\mathbb{Z}^3$ eventually returns to the origin is strictly less than 1.

We now consider two programs that require *continuous* distributions. The semantics of such programs cannot be defined without the full power of *measure theory*, the mathematical foundation of probabilities and integration. The program of Fig. 1.3 approximates the constant π using *Monte Carlo integration*, a probabilistic integration method. The program works by taking a large number of independent, uniformly distributed random samples from the square $[0,1] \times [0,1]$ and counting the number that fall inside the unit circle. As the area of the square is 1 and the area of the part of the unit circle inside that square is $\pi/4$, by the law of large numbers we expect to see a $\pi/4$ fraction of sample points lying inside the circle.

```
i:=0;
n:=0;
while i<1e9 do
   x:=rand();
   y:=rand();
   if (x*x+y*y) < 1 then n:=n+1;
   i:=i+1
i:=4*n/1e9;
```

Figure 1.3 Probabilistic computation of π.

In this example, the random number generator `rand()` samples from the uniform distribution on the interval $[0,1]$. This distribution is often called *Lebesgue measure*. Here the state space $[0,1]$ is uncountable and the probability of drawing any particular $x \in [0,1]$ is zero. Such probability distributions are called *continuous*. The natural question to ask about this program is not whether it terminates (it clearly does) but whether it returns a good approximation of π with high probability. We will answer this question in Section 1.3.

Finally, the program in Fig. 1.4 generates a real number between $[0,1]$ whose expansion in base 3 does not contain any 1's. This program is not like the others in that it does not halt (nor is it meant to). The program generates a sample from a curious and in many respects counterintuitive distribution called the *Cantor distribution*. It cannot be described using discrete probability distributions (i.e. finite or countable weighted sums of point masses), although the program only uses a discrete fair coin as a source. The Cantor distribution is also an example of *continuous* probability distribution, which assigns probability zero to every element of the state space. It is also an example of a so-called *singular* distribution, since it can be shown that the set of all its possible outcomes—that is to say the set of

all real numbers whose base-3 expansion contains no 1's—has measure 0 in the Lebesgue measure on $[0, 1]$.

```
x:=0;
d:=1;
while true do
   d:=d/3;
   x:=x+2*coin()*d
```

Figure 1.4 Cantor distribution program.

1.2 Measure theory: What you need to know

Measures are a generalization of the concepts of length, area, or volume of Euclidean geometry to other spaces. They form the basis of probability and integration theory. In this section, we explain what it means for a space to be a *measurable space*, we define *measures* on these spaces, and we examine the rich structure of *spaces of measures*, which will be essential to the semantics of probabilistic programs defined in Section 1.3.5. When not specified otherwise we use the word measure to refer to finite measures.

1.2.1 Some intuition

The concepts of length, area, and volume on Euclidean spaces are examples of *(positive) measures*. These are sufficient to illustrate most of the desired properties of measures and some pitfalls to avoid. For the sake of simplicity, let us examine the concept of *length*. Given an interval $[a, b] \subseteq \mathbb{R}$, its length is of course $\ell([a, b]) = b - a$. But the length function ℓ makes sense for other subsets of $\mathbb{R}$ besides intervals. So we will begin with two related questions:

(a) Which subsets of $\mathbb{R}$ can meaningfully be assigned a "length" consistent with the length of intervals? I.e., what should the domain of ℓ be?
(b) Which properties should the length function ℓ satisfy?

The answer to question (a) will give rise to the notion of *measurable space*, and the answer to question (b) will give rise to the notion of *measure*, both defined formally in Section 1.2.2.

Note that larger intervals have larger lengths: if $[a, b] \subseteq [c, d]$, then we have that $\ell([a, b]) = b - a \leq d - c = \ell([c, d])$. This intuitively obvious property is a general feature of all positive measures: they associate nonnegative real numbers to subsets monotonically with respect to set inclusion. Let us now take two disjoint intervals

$[a_1, b_1]$ and $[a_2, b_2]$ with $b_1 < a_2$. It is natural to define the length of $[a_1, b_1] \cup [a_2, b_2]$ as the sum of the length of the respective intervals, i.e.

$$\ell([a_1, b_1] \cup [a_2, b_2]) = \ell([a_1, b_1]) + \ell([a_2, b_2]) = (b_1 - a_1) + (b_2 - a_2).$$

We can draw two conclusions from this natural definition. First, if A, B are two disjoint subsets of $\mathbb{R}$ in the domain of ℓ, then their union should also belong to the domain of ℓ, and the measure of the union should be the sum of the measures. More generally, if A_i, $1 \leq i \leq n$, is any finite collection of pairwise disjoint sets in the domain of ℓ, then $\bigcup_{i=1}^{n} A_i$ should also be in the domain of ℓ, and the measure of the union should be the sum of the measures of the A_i; that is,

$$\ell\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \ell(A_i). \tag{1.1}$$

A real-valued function on subsets satisfying (1.1) is called (finitely) *additive*. All measures will be finitely additive, and in fact more. Consider the countable collection of pairwise disjoint intervals $[n, n + 2^{-n}), n \in \mathbb{N}$. Generalising (1.1), it is natural to define ℓ on the union of these intervals as

$$\ell\left(\bigcup_{n=0}^{\infty} [n, n + 2^{-n})\right) = \sum_{n=0}^{\infty} 2^{-n} = 2.$$

Again, we can draw two conclusions from this natural definition. First, if A_i for $i \in \mathbb{N}$ is a *countable* collection of pairwise disjoint sets in the domain of ℓ, then $\bigcup_{i \in \mathbb{N}} A_i$ should be in the domain of ℓ; second, that (1.1) should be extended to such countable collections, so that

$$\ell\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \ell(A_i). \tag{1.2}$$

A function ℓ satisfying (1.2) is called *countably additive* or *σ-additive*. Every measure will be countably additive. The reader will now legitimately ask: what happens if the sum in (1.2) diverges? To deal with this behaviour, one simply allows ∞ as a possible length, that is to say the codomain of ℓ can be the extended real line $\mathbb{R}^+ \cup \{\infty\}$. In particular, this allows us to define the length of $\mathbb{R}$ via (1.2) as:

$$\ell(\mathbb{R}) = \ell\left(\bigcup_{n \in \mathbb{Z}} [n, n + 1)\right) = \infty.$$

However, for the purpose of semantics of probabilistic programs, we will not need measures taking the value ∞. A measure is called *finite* if it only assigns finite values in $\mathbb{R}$ to any set in its domain. *For the remainder of this chapter, the term "measure", otherwise unqualified, will refer to* finite *measures.*

Consider now subsets $A \subseteq B$ of $\mathbb{R}$ in the domain of ℓ such that $\ell(A) \leq \ell(B) < \infty$. From finite additivity, it would make sense to define $\ell(B \setminus A) = \ell(B) - \ell(A)$, since $B = A \cup (B \setminus A)$ is a partition of B. In other words, it would also be natural to require that if $A \subseteq B$ and A and B are in the domain of ℓ, then so should be $B \setminus A$, and $\ell(B \setminus A) = \ell(B) - \ell(A)$. Thus the domain of ℓ should be closed under complementation.

The reader may now be wondering: If the domain of ℓ contains all intervals and is closed under countable pairwise disjoint unions and complementation, that is already a very large set of subsets of $\mathbb{R}$. Is it possible that a length can be sensibly assigned to *all* subsets of $\mathbb{R}$? In other words, can we extend ℓ to domain $\mathcal{P}(\mathbb{R})$? Alas, it turns out that this is not possible. An important and desirable property of the length function ℓ is that it is *translation invariant*: given a set A with length $\ell(A)$ (for example an interval), if the entire set A is translated a fixed distance, say d, then its length should be unchanged; that is, $\ell(A) = \ell(\{x + d \mid x \in A\})$. Vitali (1905) constructed a countable set of subsets of the interval $[0, 1)$, called *Vitali sets*, which are pairwise disjoint, translates of each other (modulo 1), and whose union is $[0, 1)$. They would all have to have the same measure, which would break the countable additivity axiom (1.2). Vitali sets are examples of *non-measurable sets*. They provide an example of subsets of $\mathbb{R}$ which are incompatible with the basic assumptions of how the length function should behave. Thus the domain of the length function cannot be $\mathcal{P}(\mathbb{R})$, because it cannot contain the Vitali sets.

The length function ℓ described in the preceding paragraphs is called the *Lebesgue measure* on $\mathbb{R}$. We now turn our attention to axiomatizing the intuitive ideas presented thus far.

1.2.2 Measurable spaces and measures

We start by axiomatizing the closure properties of the domain of a measure (such as the length function) which we have described informally in the previous section.

A *σ-algebra* $\mathcal{B}$ on a set S is a collection of subsets of S containing the empty set $\varnothing$ and closed under complementation in S and countable union (hence also under countable intersection). A pair $(S, \mathcal{B})$, where S is a set and $\mathcal{B}$ is a σ-algebra on S, is called a *measurable space*. The elements of $\mathcal{B}$ are called the *measurable sets* of the space. In a probabilistic setting, elements of S and $\mathcal{B}$ are often called *outcomes* and *events*, respectively. The domain of a measure, for example the length function, will always be a σ-algebra. If the σ-algebra is obvious from the context, we simply say that S is a measurable space. The set of all subsets $\mathcal{P}(S)$ is a σ-algebra called the *discrete σ-algebra*, but as noted above, it may not be an appropriate choice since it may not allow the definition of certain measures. However, it is always an

acceptable choice for finite or countable sets, and we will always assume that finite and countable sets are equipped with the discrete σ-algebra.

If $\mathcal{F}$ is a collection of subsets of a set S, we define $\sigma(\mathcal{F})$, the σ-algebra *generated* by $\mathcal{F}$, to be the smallest σ-algebra containing $\mathcal{F}$. That is, $\sigma(\mathcal{F})$ is the smallest collection of subsets of S containing $\mathcal{F}$ and $\varnothing$ and closed under countable union and complement. Equivalently,

$$\sigma(\mathcal{F}) \triangleq \bigcap\{\mathcal{A} \mid \mathcal{F} \subseteq \mathcal{A} \text{ and } \mathcal{A} \text{ is a } \sigma\text{-algebra}\}.$$

Note that $\sigma(\mathcal{F})$ is well-defined, since the intersection is nonempty, as $\mathcal{F} \subseteq \mathcal{P}(S)$ and $\mathcal{P}(S)$ is a σ-algebra. If $(S, \mathcal{B})$ is a measurable space and $\mathcal{B} = \sigma(\mathcal{F})$, we say that the space is *generated* by $\mathcal{F}$.

Measurable functions. Let $(S, \mathcal{B}_S)$ and $(T, \mathcal{B}_T)$ be measurable spaces. A function $f\colon S \to T$ is *measurable* if the inverse image $f^{-1}(B) = \{x \in S \mid f(x) \in B\}$ of every measurable subset $B \in \mathcal{B}_T$ is a measurable subset of S. When $\mathcal{B}_T$ is generated by $\mathcal{F}$, then f is measurable if and only if $f^{-1}(B)$ is measurable for every $B \in \mathcal{F}$.

An example of a measurable function is $\chi_B\colon S \to \{0,1\}$, the *characteristic function* of a measurable set B:

$$\chi_B(s) = \begin{cases} 1, & s \in B, \\ 0, & s \notin B. \end{cases}$$

Here, $(S, \mathcal{B})$ is a measurable space, $B \in \mathcal{B}$, and $\{0,1\}$ is the discrete space.

Measures. A *signed (finite) measure* on $(S, \mathcal{B})$ is a countably additive map $\mu\colon \mathcal{B} \to \mathbb{R}$ such that $\mu(\varnothing) = 0$. Recall that *countably additive* means that if $\mathcal{A}$ is a countable set of pairwise disjoint events, then $\mu(\bigcup \mathcal{A}) = \sum_{A \in \mathcal{A}} \mu(A)$. Equivalently, if $A_0, A_1, A_2, \ldots$ is a countable chain of events (a countable collection of measurable sets such that $A_n \subseteq A_{n+1}$ for all $n \geq 0$), then $\lim_n \mu(A_n)$ exists and is equal to $\mu(\bigcup_n A_n)$.

A signed measure on $(S, \mathcal{B})$ is called *positive* if $\mu(A) \geq 0$ for all $A \in \mathcal{B}$. A positive measure on $(S, \mathcal{B})$ is called a *probability measure* if $\mu(S) = 1$ and a *subprobability measure* if $\mu(S) \leq 1$. A measurable set B such that $\mu(B) = 0$ is called a *μ-nullset*, or simply a *nullset* if there is no ambiguity. A property is said to hold *μ-almost surely* (μ-a.s.) or *μ-almost everywhere* (μ-a.e.) if the set of points on which it does *not* hold is contained in a nullset.

In probability theory, measures are sometimes called *distributions*. We will use the terms *measure* and *distribution* synonymously.

For $s \in S$, the *Dirac measure*, or *Dirac delta*, or *point mass* on s is the probability

measure

$$\delta_s(B) = \begin{cases} 1, & s \in B, \\ 0, & s \notin B. \end{cases}$$

A measure is *discrete* if it is a countable weighted sum of Dirac measures. In particular a convex sum of Dirac measures is a discrete probability measure. These are finite or countable sums of the form $\sum_{s \in C} a_s \delta_s$, where all $a_s \geq 0$ and $\sum_{s \in C} a_s = 1$.

A measure μ on a measurable set $(S, \mathcal{B})$ is called *continuous* if $\mu(\{s\}) = 0$ for all singleton sets $\{s\}$ in $\mathcal{B}$. The Lebesgue measures on $\mathbb{R}^n$ for $n \in \mathbb{N}$, that is, the lengths, areas, volumes, etc., are the best known examples of continuous measures.

Pushforward measure. Given $f : (S, \mathcal{B}_S) \to (T, \mathcal{B}_T)$ measurable and a measure μ on $\mathcal{B}_S$, one defines the *pushforward measure* $f_*(\mu)$ on $\mathcal{B}_T$ by

$$f_*(\mu)(B) = \mu(f^{-1}(B)), \quad B \in \mathcal{B}_T. \tag{1.3}$$

This measure is well defined: since f is measurable, f^{-1} maps measurable sets of $\mathcal{B}_T$ to measurable sets of $\mathcal{B}_S$.

Lebesgue integration. An important operation on measures and measurable functions is *Lebesgue integration*. Let $(S, \mathcal{B})$ be a measurable space. Given a measure $\mu : \mathcal{B} \to \mathbb{R}$ and bounded measurable function $f : S \to \mathbb{R}$, say bounded above by M and below by m, the *Lebesgue integral* of f with respect to μ, denoted $\int f \, d\mu$, is a real number obtained as the limit of finite weighted sums of the form

$$\sum_{i=0}^{n} f(s_i)\mu(B_i), \tag{1.4}$$

where $B_0, \ldots, B_n$ is a measurable partition of S, the value of f does not vary more than $(M-m)/n$ in any B_i, and $s_i \in B_i$, $1 \leq i \leq n$. The limit is taken over increasingly finer measurable partitions of the space. For the details of this construction, see for example (Dudley, 2002, Ch. 4) or (Aliprantis and Border, 1999, Ch. 11).

For a finite discrete space $n = \{1, 2, \ldots, n\}$, the integral reduces simply to a weighted sum: $\int f \, d\mu = \sum_{i=1}^{n} f(i)\mu(i)$.

The *bounded integral* $\int_B f \, d\mu$, where $B \in \mathcal{B}$, is obtained by integrating over the set B instead of all of S; equivalently,

$$\int_B f \, d\mu \triangleq \int \chi_B \cdot f \, d\mu, \tag{1.5}$$

where χ_B is the characteristic function of B and $\chi_B \cdot f$ is the pointwise product of real-valued functions.

Absolute continuity. Given two measures μ, ν, we say that μ is *absolutely continuous* with respect to ν and write $\mu \ll \nu$ if for all measurable sets B, if $\nu(B) = 0$, then $\mu(B) = 0$. Informally, if ν assigns no mass to B, then neither does μ. Although we will not need it, we cannot fail to mention the following theorem, which is one the pillars of probability theory.

Theorem 1.1 (Radon–Nikodym) *Let μ, ν be two finite measures on a measurable space $(S, \mathcal{B})$ and assume that μ is absolutely continuous with respect to ν. Then there exists a measurable function $f : S \to \mathbb{R}$ defined uniquely up to a μ-nullset such that*

$$\mu(B) = \int_B f \, d\nu.$$

The function f is called the Radon–Nikodym derivative *of μ with respect to ν.*

Radon–Nikodym derivatives are known in probability theory as *probability density functions*. For example, the standard Gaussian probability measure is absolutely continuous with respect to Lebesgue measure (the length function) on $\mathbb{R}$. Its Radon–Nikodym derivative with respect to Lebesgue measure is the Gaussian density function $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Products. Given two measurable spaces $(S_1, \mathcal{B}_1)$ and $(S_2, \mathcal{B}_2)$, one can construct the *product space* $(S_1 \times S_2, \mathcal{B}_1 \otimes \mathcal{B}_2)$, where $S_1 \times S_2$ is the cartesian product and $\mathcal{B}_1 \otimes \mathcal{B}_2$ is the σ-algebra on $S_1 \times S_2$ generated by all *measurable rectangles* $B_1 \times B_2$ for $B_1 \in \mathcal{B}_1$ and $B_2 \in \mathcal{B}_2$. In other words,

$$\mathcal{B}_1 \otimes \mathcal{B}_2 \triangleq \sigma\left(\{B_1 \times B_2 \mid B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}\right). \tag{1.6}$$

The measurable rectangles $B_1 \times B_2$ are a generalisation of the case where $S_1 = S_2 = \mathbb{R}$ and B_1, B_2 are intervals. The product of two measurable spaces is thus the measurable space generated by the corresponding measurable rectangles.

A measure on the product space $(S_1 \times S_2, \mathcal{B}_1 \otimes \mathcal{B}_2)$ is sometimes called a *joint distribution*. Due to the inductive construction (1.6) of $\mathcal{B}_1 \otimes \mathcal{B}_2$ from measurable rectangles $B_1 \times B_2$, joint distributions are uniquely determined by their values on measurable rectangles. For details of this extension, see (Dudley, 2002, §4.4).

A special class of joint distributions are the *product measures* $\mu_1 \otimes \mu_2$ formed from a measure μ_1 on $(S_1, \mathcal{B}_1)$ and a measure μ_2 on $(S_2, \mathcal{B}_2)$, defined on measurable rectangles by

$$(\mu_1 \otimes \mu_2)(B_1 \times B_2) \triangleq \mu_1(B_1)\mu_2(B_2).$$

As mentioned, this extends uniquely to a joint distribution $\mu_1 \otimes \mu_2 : \mathcal{B}_1 \otimes \mathcal{B}_2 \to \mathbb{R}$. Product measures capture the idea of *independence*: sampling $\mu_1 \otimes \mu_2$ to obtain an element of $S_1 \times S_2$ is equivalent to independently sampling μ_1 on S_1 and μ_2 on S_2.