Introduction to Probability Theory: A Measure Theoretic Viewpoint

July 2020

Balaram D. Behera
bbehera@ucsc.edu
Mathematics 105A
Real Analysis
University of California
Santa Cruz

Abstract

Measures are everywhere: we have the counting measure which signifies the size of sets, the Dirac measure which signifies if an element is contained in a set, and of course the Lebesgue measure. Furthermore, the probability function is another measure, and in this paper we will introduce probability theory in the form of measures. Before diving into probability theory, we will spend a good amount of time building the basics of measure theory by first introducing σ-algebras, the Lebesgue integral, and then move on to some important convergence results. Finally, we start defining the fundamentals of probability theory like random variables, distributions, density, and expectation in terms of probability measure spaces. In conclusion, the goal of this paper is to provide a fresh viewpoint of probability theory in the rich context of measure theory.

1 Introduction

As mentioned above, measures are everywhere and it's how we navigate our way through real numbers. In Theoretical Computer Science, we tend to work in the discrete sense, but in recent decades, that has evolved into simultaneously working in the continuous space, especially with optimization theory and probability theory. Most probability theory in Theoretical Computer Science deals with countable sample spaces, but at the same time, the theory at its root deals with continuous sample spaces; and with Machine Learning, the continuous space is used (of course up to the computable precision). This stresses the importance of taking a look at probability in the continuous sense, and of course this can be represented as a measure. This is the end goal of this paper — to provide this measure theoretic viewpoint.

We will first begin with the fundamentals of measure theory, by first talking about σ-algebra and the Borel σ-algebra. Then we proceed to discuss measurable spaces, measures, and measure spaces. Then we discuss the Lebesgue measure which attempts at solving at the unsolvable measure problem. From there we immediately delve into the Lebesgue integral with a brief motivation for its definition. We then discuss very briefly the difference between this integral and the more popular Riemann approach. From there we go into the Monotone Convergence Theorem, Fatou's Theorem, and the Dominated Convergence Theorem which are the foundational "big" theorems of measure theory and integration theory. Penultimately, we explore the beginnings of probability theory with the traditional approach, and then, lastly, discuss our measure theoretic approach. All these definitions and theorems are adapted from [10] and [3].

2 Sigma Algebra

Let's first define σ-algebras which is the basis definition of measure theory (other than the set, of course). Recall that $\mathcal{P}(X)$ denotes the power set of the set X.

Definition 1. *For a set X, $\mathcal{A} \subseteq \mathcal{P}(X)$ is a σ-algebra if*

(a) $\varnothing \in \mathcal{A}$ *and* $X \in \mathcal{A}$

(b) $A \in \mathcal{A}$ *implies that* $\overline{A} = X \setminus A \in \mathcal{A}$

(c) $A_i \in \mathcal{A}$ *for* $i \in \mathbb{N}$ *implies that* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Thus we see that a σ-algebra is closed under the complement and union operations on all its elements. Now we define what the elements of a σ-algebra are called.

Definition 2. *For a σ-algebra $\mathcal{A} \subseteq \mathcal{P}(X)$ for some set X, any $A \in \mathcal{A}$ is called a measurable set (also denoted as an $\mathcal{A}$-measurable set).*

So we see that all sets in a σ-algebra are what we call measurable, and this notion will be become relevant and surface once we define a measure.

Notice that according to the requirements for being a σ-algebra, the minimal σ-algebra $\mathcal{A} \subseteq \mathcal{P}(X)$ is $\mathcal{A} = \{\varnothing, \mathcal{P}(X)\}$ as we are closed under the complement and union. Now let's take a look at this intriguing notion of a minimal σ-algebra.

Definition 3. *For some $M \subseteq \mathcal{P}(X)$, there is a minimal σ-algebra that contains M defined as,*

$$\sigma(M) = \bigcap_{\mathcal{A} \supseteq M, \mathcal{A} \ \sigma\text{-algebra}} \mathcal{A}$$

which is also known as the σ-algebra generated by M.

Now in the special case where X is a metric space, i.e. $X = \mathbb{R}^n$ for some $n \in \mathbb{Z}^+$, we have the following definition.

Definition 4. *The σ-algebra generated by $\mathbb{R}^n$ is called the Borel σ-algebra and is denoted by $\mathcal{B}(\mathbb{R}^n)$ (also denoted as $\mathcal{B}_n$), and each $B \in \mathcal{B}_n$ is called a Borel set or a Borel measurable set.*

The Lebesgue measure deals with metric spaces primarily, so we will be using this Borel σ-algebra all throughout this paper. Now that we have the foundation of σ-algebras set, let's continue into defining a measure.

3 The Measure Problem

Let's first define this relation between a set X and some σ-algebra $\mathcal{A} \subseteq \mathcal{P}(X)$.

Definition 5. *The pair $(X, \mathcal{A})$ is called a measurable space if $\mathcal{A} \subseteq \mathcal{P}(X)$ is a σ-algebra.*

Now we are ready to define a measure.

Definition 6. *Let $(X, \mathcal{A})$ be a measurable space. A map $\mu : \mathcal{A} \to [0, \infty]$ is called a measure if*

(a) $\mu(\varnothing) = 0$

(b) $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ *for* $A_i \cap A_j = \varnothing$ *for all* $i \neq j$*, and each* $A_i \in \mathcal{A}$.

The second property is called σ-additivity and is handy since now we are able to satisfy the intuitive idea that measures should add up. Also note that a measure can map certain measurable sets to ∞. Lastly, let's define the measure space as follows.

Definition 7. *The triple $(X, \mathcal{A}, \mu)$ is called a measure space if $(X, \mathcal{A})$ is a measurable space and $\mu : \mathcal{A} \to [0, \infty]$ is a measure.*

To provide an example, consider the Counting measure defined as follows on some σ-algebra $\mathcal{A}$:

$$\mu(A) = \begin{cases} |A| & \text{if } A \text{ has finitely many elements} \\ \infty & \text{otherwise} \end{cases}.$$

Here we have $\mu : \mathcal{A} \to [0, \infty]$ by definition of the cardinality function on finite sets, and also $\mu(\varnothing) = 0$ obviously. For mutually disjoint $A_1, \ldots, A_n \in \mathcal{A}$, it's obvious that we have

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i),$$

but in the infinite case, since $c \cdot \infty = \infty$ for $c \in \mathbb{N}$, we satisfy the σ-additive property of a measure as well with μ.

Now going back to the real numbers, we want to find the ideal measure μ over the measurable space $(\mathbb{R}^n, \mathcal{A})$, i.e. $\mu([0,1]^n) = 1$ (the unit volume) and $\mu(x + A) = \mu(A)$ for $x \in \mathbb{R}^n$ and $A \in \mathcal{A}$ (the measure is "location" invariant). In fact this is the definition of the Lebesgue measure. You may wonder why we left $\mathcal{A}$ as arbitrary because ideally we want $\mathcal{A} = \mathcal{P}(\mathbb{R}^n)$, i.e. all subsets of $\mathbb{R}^n$ are measurable, but actually this is not true. There exists no μ over the measurable space $(\mathbb{R}^n, \mathcal{P}(\mathbb{R}^n))$ such that we satisfy the conditions of the Lebesgue measure. Notice that the Lebesgue measure μ in represents the traditional length, area, and volume that we know of in the metric spaces $\mathbb{R}$, $\mathbb{R}^2$, and $\mathbb{R}^3$, respectively. Please note that I have omitted the lengthy derivation of the Lebesgue measure through the product measure because the intuitive definition is sufficient for the understanding of this paper; however, we did spend time on the abstract measure theory topics to show the similarity in the definition of probability spaces in a future section.

4 The Lebesgue Integral

Having spent a considerable amount of time building up the foundational aspects of measure theory and introducing the Lebesgue measure, let's now dive into the heart of all this: the Lebesgue integral (not to be confused with the more commonly known Reimann integral).

But before talking about integrals, we must first introduce the notion of measurable maps.

Definition 8. *Let* $(\Omega_1, \mathcal{A}_1)$ *and* $(\Omega_2, \mathcal{A}_2)$ *be two measurable spaces. Then the function* $f : \Omega_1 \to \Omega_2$ *is* <u>*measurable*</u> *(with respect to* $\mathcal{A}_1$ *and* $\mathcal{A}_2$*) if* $f^{-1}(A) \in \mathcal{A}_1$ *for all* $A \in \mathcal{A}_2$*; in other words, every pre-image of an* $\mathcal{A}_2$*-measurable set is also an* $\mathcal{A}_1$*-measurable set.*

This notion is important once we start looking at functions in the light of measurable maps to aide in our Lebesgue integral definition.

Let's first provide an example of a measurable map: consider the measurable spaces $(\Omega, \mathcal{A})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and the characteristic function $\chi_A : \Omega \to \mathbb{R}$ for $A \in \mathcal{A}$ defined by

$$\chi_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}.$$

We see that for all $\mathcal{A}$-measurable sets A, the function χ_A is a measurable map since

$$\chi_A^{-1}(\varnothing) = \varnothing, \quad \chi_A^{-1}(\mathbb{R}) = \Omega, \quad \chi_A^{-1}(\{1\}) = A, \quad \chi_A^{-1}(\{0\}) = \overline{A}$$

and since $A \in \mathcal{A}$ and since $\mathcal{A}$ is a σ-algebra we know that $\{\varnothing, \Omega, A, \overline{A}\} \subseteq \mathcal{A}$ as required.

Now let's state a few a nice properties of measurable maps.

Proposition 1. *Let* $(\Omega_1, \mathcal{A}_1)$*,* $(\Omega_2, \mathcal{A}_2)$*, and* $(\Omega_3, \mathcal{A}_3)$ *be measurable spaces, and let* $f : \Omega_1 \to \Omega_2$ *and* $g : \Omega_2 \to \Omega_3$ *be measurable maps. Then* $f \circ g$ *is also a measurable map.*

Proposition 2. *Consider the measurable spaces* $(\Omega, \mathcal{A})$ *and* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$*, and let* $f, g : \Omega \to \mathbb{R}$ *be measurable maps. Then the functions* $f + g$*,* $f - g$*,* $f \circ g$*, and* $|f|$ *are also measurable.*

Especially, the second proposition is key to understanding the applicability of measure theory in "divisions" analogous to how we have Reimann Sums.

Finally we have the sufficient material to begin our shallow study of the all-important Lebesgue integral. First we will define it for simple functions (also known as step functions), and that will suffice to construct a more general definition of the integral.

Let's first define the notion of a simple function.

Definition 9. *Consider some measure space* $(X, \mathcal{A}, \mu)$*. Then any function* $f : X \to \mathbb{R}$ *that can be defined by*

$$f(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(A_i),$$

where $A_1, \ldots, A_n \in \mathcal{A}$ measurable sets and $c_1, \ldots, c_n \in \mathbb{R}$ and where χ_{A_i} is the characteristic function of the set $A_i \subseteq X$, is called a simple function.

Further, let the set of all simple functions be denoted by $\mathcal{S}$. Now from this, we clearly see that for some $f \in \mathcal{S}$ defined with only non-negative c_i, we get (where $I(.)$ is the Lebesgue integral function with respect to μ):

$$I(f) = \sum_{i=1}^{n} c_i \mu(A_i)$$

just with the basic intuitive idea that the integral is the "area under the curve." However, the assumption that all c_i are non-negative is crucial because that is the only we avoid a "$\infty - \infty = ?$" situation for this. We call this type of function a non-negative simple function, and the set of all such functions is denoted by $\mathcal{S}^+ \subseteq \mathcal{S}$ and formally defined as

$$\mathcal{S}^+ = \{f : X \to \mathbb{R} \mid f \text{ simple function}, f \geq 0\}.$$

To sum up our conclusion regarding Lebesgue integrals on positive simple functions, (under the measure space $(X, \mathcal{A}, \mu)$) consider a function $f \in \mathcal{S}^+$ with representation

$$f(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(A_i)$$

where $c_i \geq 0$ for all i. Then the Lebesgue integral of f with respect to our measure μ (the Lebesgue measure), we have

$$\int_X f(x)\, d\mu(x) = \int_X f\, d\mu = I(f) = \sum_{i=1}^{n} c_i \mu(A_i) \in [0, \infty]$$

which is well-defined. It's also important to note that I is monotonically increasing, i.e. if functions $f, g \in \mathcal{S}^+$ and $f \leq g$, we have $I(f) \leq I(g)$.

Now we are ready to generalize the Lebesgue integral given in the following definition.

Definition 10. *Consider the measure space $(X, \mathcal{A}, \mu)$ and some measurable function $f : X \to [0, \infty)$. Then the Lebesgue integral of f with respect to μ is given by*

$$\int_X d\, d\mu = \sup\{I(h) \mid h \in \mathcal{S}^+, h \leq f\}.$$

We call f μ-integrable (for the Lebesgue measure we say Lebesgue-integrable if $\int_X f\, d\mu \leq \infty$.

Essentially, we look for the best step function that represents our function f, and then realize that the integral of the step function is equal to the integral of the function and this is accomplished by the use of the supremum.

Now that we have the Lebesgue integral it begs to answer the question: "How is this different from the Riemann integral?" Firstly, Reimann integrals are hard to transfer to higher dimensions since the idea of "vertical cuts" doesn't transpire well in higher dimensions; also, it depends on continuous functions; and we don't have certain convergence theorems that the Lebesgue integral provides. However, it's important to note that the Lebesgue integral requires the concept of a measure, and in metric spaces the Lebesgue measure suffices. I only provide a brief description of their differences and the reason why the Lebesgue integral is more general because our goal is to go in the direction of applying this to probability theory.

5 The Convergence Theorems

Although this topic is a short digression from our end goal, I believe it's imperative to include the convergence theorems in an introduction to measure theory. We will simply state the theorems, namely, the Monotone Convergence Theorem, Fatou's Lemma, and Lebesgue's Dominant Convegrence Theorem, and their brief explanations.

For this section, assume that we have the measure space $(X, \mathcal{A}, \mu)$. And, before continuing on, we need to first define a new term that arises with the advent of measures.

Definition 11. *A property P is said to hold almost everywhere with respect to μ (equivalently denoted as μ-a.e.) if there exists a set $N \in \mathcal{A}$ with $\mu(N) = 0$ such that all $x \in X \setminus N$ satisfy property P.*

Now we can state the following properties we have with measurable functions.

Proposition 3. *Let $f, g : X \to [0, \infty)$ be measurable functions. Then the following hold:*

(a) if $f = g$ μ-a.e., then $\int_X f\, d\mu = \int_X g\, d\mu$;

(b) if $f \leq g$ μ-a.e., then $\int_X f\, d\mu \leq \int_X g\, d\mu$;

(c) and $f = 0$ μ-a.e. if and only if $\int_X f\, d\mu = 0$.

Essentially, these properties show that the Lebesgue integral ignores differences in functions that have measure zero, and this is quite important to note as these properties aide in the derivation of the convergence theorems.

Now let's take a look at the important Monotone Convergence Theorem.

Theorem 4. *Let $f_n : X \to [0, \infty)$ be a measurable function for all $n \in \mathbb{N}$ such that μ-a.e. we have $f_1 \leq f_2 \leq \cdots$. Then μ-a.e. ($x \in X$) we have,*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X \left(\lim_{n\to\infty} f_n \right) \, d\mu.$$

This property is held for all monotonic sequences of functions for the Lebesgue integral, but for the Riemann integral, the types of functions are much more specific, and this convergence theorem adds to the generality and usability of the Lebesgue integral.

In buildup to the next very crucial and fundamental convergence theorem, we have Fatou's Lemma below which isn't an extremely strong claim but still very helpful.

Lemma 5. *Let $f_n : X \to [X, \infty]$ be a measurable function for all $n \in \mathbb{N}$. Then μ-a.e. ($x \in X$) we have,*

$$\int_X \left(\liminf_{n\to\infty} f_n \right) \, d\mu \leq \liminf_{n\to\infty} \int_X f_n \, d\mu.$$

The proof of this lemma follows from the definition of the limit inferior and the use of the Monotone Convergence Theorem. Now for the next theorem, we need the following set:

$$\mathcal{L}(\mu) = \left\{ f : X \to \mathbb{R} \text{ measurable} \,\middle|\, \int_X |f| \, d\mu < \infty \right\}.$$

Notice that $\mathcal{L}(\mu)$ is the set of all Lebesgue-integrable functions. Also note that for any $f \in \mathcal{L}(\mu)$ we can write $f = f^+ - f^-$ for $f^+, f^- \geq 0$, and we need this because we defined Lebesgue integration only on non-negative functions. Furthermore, we define

$$\int_X f \, d\mu = \int_X f^+ \, d\mu - \int_X f^- \, d\mu.$$

Now we can state Lebesgue's Dominated Convergence Theorem as follows.

Theorem 6. *Let $f_n : X \to \mathbb{R}$ be a measurable function for all $n \in \mathbb{N}$, and let $g \in \mathcal{L}(\mu)$ such that $|f_n| \leq g$ for all $n \in \mathbb{N}$ (we call g an integrable majorant).*

Also let $f : X \to \mathbb{R}$ such that $f = \lim_{n\to\infty} f_n$ for all $x \in X$ μ-a.e. Then we know that $f_1, f_2, \ldots \in \mathcal{L}(\mu)$ and $f \in \mathcal{L}(\mu)$, and we have

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu.$$

This result is very big for integration theory, in particular, but I mention it only for the sake of writing a comprehensive overview; we won't be using these convergence results in our development of probability theory.

6 Measure Theoretic Probability Theory

Finally, after our extensive yet quick overview of some of the basics of measure theory, we are ready to start applying the same concepts to develop probability theory from scratch. We will first start by introducing the probability spaces, and then discuss events, probability, and independence. Then we will discuss random variables, their probabilities, distribution, density and expectation, and with that we will complete our short introduction to probability theory.

First let's define the probability space which is really a measure space, and this space will include the probability measure which is the basis of probability theory.

Definition 12. *A probability space is a measure space $(\Omega, \mathcal{F}, \Pr)$ where*

(a) $\Omega \neq \varnothing$ is the set of all samples or the sample space;

(b) $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra and $E \in \mathcal{F}$ are called events;

(c) and $\Pr$ is the probability measure on $\mathcal{F}$ where $\Pr(\Omega) = 1$ and also note that $\Pr$ is a finite measure; for each event $E \in \mathcal{F}$, $\mathcal{P}(E)$ is the probability of E.

Notice that we only have the constraint that the probability of the sample space is 1. The other traditional constraint that $\Pr : \Omega \to [0, 1] \subseteq \mathbb{R}$ is also satisfied by the σ-additivity of a measure even though we don't explicitly define it as such. For the rest of this section assume we are given a probability space $(\Omega, \mathcal{F}, \Pr)$.

Now let's move on to the significant topic of independence which is the desired state we wish to have but is extremely Utopian in the sense that reality is full of dependence.

Definition 13. *Two events $A, B \in \mathcal{F}$ are independent if*

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

This definition is very primitive for a reason, and with more high-level derivations we can derive more elaborate forms of characterizing independence; but to preserve the length of this paper, we don't describe such results here.

Moving on to the concept of the random variable which is key to all probabilistic analysis and randomized algorithm analysis, let's first define this notion in accordance with our probability space $(\Omega, \mathcal{F}, \Pr)$.

Definition 14. *Any measurable function $X : \Omega \to \mathbb{R}$ is called a random variable. Further, for $x \in \mathbb{R}$, we define $\Pr(X \leq x)$ as the probability of the event $\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$.*

Note that the definition of a random variable is extremely abstract which allows us to have all kinds of distributions of random variables as we will discuss later. More importantly, the notion of $\Pr(X \leq x)$ essentially refers to the probability that the random variable X is bounded by x, and all such possibilities are captured by the event $\{\omega \in \Omega \mid X(\omega) \leq x\}$ which is in $\mathcal{F}$ since X is measurable (and that set is a union of pre-images).

Now let's talk about the concept of a random variable's distribution, and beware that the definition is quite abstract and non-intuitive initially; in fact, the distribution of a random variable is actually another measure. For random variable X, let's first define the function $\Pr_X : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ as follows for some $A \in \mathcal{B}(\mathbb{R})$:

$$\Pr_X(A) = \Pr(X^{-1}(A)).$$

We know for every random variable X, the function $\Pr_X$ is a probability measure on $\mathcal{B}(\mathbb{R})$, and for any probability measure μ there exists a random variable X such that $\Pr_X = \mu$. Then we define the distribution of X by the probability measure $\Pr_X$ on the real Borel σ-algebra; equivalently, $\Pr_X(A)$ of a $A \in \mathcal{B}(\mathbb{R})$ is the probability that the random variable X will be contained in A. Further, any probability measure on $\mathcal{B}(\mathbb{R})$ itself is a distribution of some random variable. Now moving on to density of a random variable, we would have

$$\Pr_X(A) = \int_A f \, d\Pr$$

and that function f is called the density of the distribution $\Pr_X$.

Lastly, let's define expectation as follows.

Definition 15. *Let X be a random variable. It's* <u>*expectation*</u> *is defined by*

$$\mathbb{E}(X) = \int_{\Omega} X(\omega)\, d\Pr(\omega).$$

The important properties of the expectation such as <u>Linearity of Expectations</u> follows from this definition as well. Furthermore, with this definition we can expand into defining <u>variance</u>, etc. However, we conclude our study of probability theory here with only the preliminary definitions.

7 Conclusion

In this paper, we explored the fundamentals of measure theory abstractly, and then we focused on the Lebesgue measure and Lebesgue integral and developed measure theory sufficiently. We even included the convergence theorems for the completeness of this study. Finally, in the last section we studied probability theory in the eyes of measures: in particular, the probability measure. We lastly studied the preliminaries in this context, especially the basics of random variables. I aimed at giving a comprehensive, yet brief, synopsis of this unassuming intersection of measure theory and probability theory which is impactful in Theoretical Computer Science as mentioned at the beginning. I, personally, find this intersection of the two theories fascinating, and unfortunately we weren't able to get very far with the probability theory parts, but still the learning has been very worthwhile. The hope is that the intrigue transfers to you, the reader, and it's my wish that this paper inspires you to continue the study of probability theory in this light of measure theory.

References

[1] *Almost everywhere*. Mar. 2020. URL: `https://en.wikipedia.org/wiki/Almost_everywhere`.

[2] William Feller. *An Introduction to Probability Theory and Its Applications: Vol. 1*. John Wiley & Sons, 1970.

[3] Alexander Grigoriyan. *Measure theory and probability*. 2007-2008. URL: `https://www.math.uni-bielefeld.de/~grigor/mwlect.pdf`.

[4] Michael Mitzenmacher and Eli Upfal. *Probability and computing: randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

[5] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[6] *Simple function*. June 2020. URL: https://en.wikipedia.org/wiki/Simple_function.

[7] Terence Tao. *An introduction to measure theory*. American Mathematical Society, 2011.

[8] Terence Tao. *Analysis I*. Hindustan Book Agency, 2017.

[9] Terence Tao. *Analysis II*. Hindustan Book Agency, 2017.

[10] Julian at The Bright Side Of Mathematics. *Measure Theory*. URL: https://www.youtube.com/playlist?list=PLBh2i93oe2qvMVqAzsX1Kuv6-4fjazZ8j.