Measure and Probability Theory

July 28, 2017

1 About

This document is part of a series of mathematical notes available at https://gwthomas.github.io/math4ml. You are free to distribute it as you wish. Please report any mistakes to gwthomas@berkeley.edu.

Measure theory is concerned with the problem of assigning a mathematically consistent notion of size to sets. We care about measure theory because of its use in the modern, rigorous formulation of probability given by Kolmogorov.

2 Collections of sets

We would like to assign measures to various subsets of $\mathbb{R}^n$ characterizing their size. Ideally our measure μ would satisfy

(i) For any countable collection of disjoint sets $E_1, E_2, \cdots \subseteq \mathbb{R}^n$,

$$\mu\left(\bigcup_i E_i\right) = \sum_i \mu(E_i)$$

(ii) If two sets $E, F \subseteq \mathbb{R}^n$ are such that E can be transformed into F by rigid transformations, then $\mu(E) = \mu(F)$.

(iii) The measure of the unit cube is 1.

The first property, called **countable additivity**, just means that if you partition a set into countably many parts, the sum of the parts' measures equals the original set's measure. The requirement that additivity hold for countable collections (as opposed to just finite collections) is important for proving limit theorems.

Unfortunately, one can show that these three properties are incompatible if we allow arbitrary subsets of $\mathbb{R}^n$. The solution in measure theory is to restrict ourselves to some "reasonable" collection of subsets.

Recall that the **powerset** of a set Ω is the set of all subsets of Ω, i.e.

$$\mathcal{P}(\Omega) = \{S : S \subseteq \Omega\}$$

Note that in particular $\varnothing, \Omega \in \mathcal{P}(\Omega)$ for any set Ω.

In the remainder we will consider collections of subsets of Ω; in other words, these collections are subsets of $\mathcal{P}(\Omega)$. We will make certain requirements of these collections so that we have some structure to work with. In particular, we choose the collections so that the properties above hold, not for arbitrary subsets of Ω, but for any sets in the collection.

2.1 Algebras and σ-algebras

Let Ω be a non-empty set. Then $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is an algebra on Ω if

(i) $\mathcal{A}$ is non-empty.

(ii) If $E \in \mathcal{A}$, then $E^c = \Omega \setminus E \in \mathcal{A}$.

(iii) If $E_1, \ldots, E_n \in \mathcal{A}$, then $\bigcup_{i=1}^n E_i \in \mathcal{A}$.

The second property states that $\mathcal{A}$ is **closed under complements**. Using de Morgan's laws, properties 2 and 3 collectively imply that $\mathcal{A}$ is closed under finite intersections as well, since

$$\bigcap_{i=1}^n E_i = \left(\bigcup_{i=1}^n E_i^c \right)^c$$

Then we must have $\varnothing \in \mathcal{A}$; since $\mathcal{A}$ is non-empty there exists some $E \in \mathcal{A}$, so $E^c \in \mathcal{A}$, and hence $\varnothing = E \cap E^c \in \mathcal{A}$.

In light of the desirability of countable additivity, we would like the collection of subsets we consider to be closed under unions of countably many sets, not just finitely many. Thus we need to strengthen condition 3, and arrive at the following definition: a **σ-algebra** is an algebra that is closed under countable unions. It follows by the same reasoning as above that a σ-algebra is also closed under countable intersections.

Note that $\{\varnothing, \Omega\}$ and $\mathcal{P}(\Omega)$ are σ-algebras for any Ω, and moreover they are respectively the smallest and largest possible σ-algebras.

If $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ is any collection of subsets of Ω, there exists a unique smallest σ-algebra containing $\mathcal{C}$; this is called the **σ-algebra generated by $\mathcal{C}$** and denoted $\sigma(\mathcal{C})$.

3 Measures

Let Ω be a non-empty set and $\mathcal{M} \subseteq \mathcal{P}(\Omega)$ a σ-algebra. The pair $(\Omega, \mathcal{M})$ is called a **measurable space**, and the elements of $\mathcal{M}$ are its **measurable sets**. A **measure** on $(\Omega, \mathcal{M})$ is a function $\mu : \mathcal{M} \to [0, \infty]$ such that

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

(ii) For any countable collection of disjoint sets $\{E_i\} \subseteq \mathcal{M}$,

$$\mu\left(\dot{\bigcup_i} E_i \right) = \sum_i \mu(E_i)$$

The triple $(\Omega, \mathcal{M}, \mu)$ is called a **measure space**.

The simplest nontrivial example of a measure is the **counting measure**, given by

$$E \mapsto \begin{cases} |E| & E \text{ finite} \\ \infty & \text{otherwise} \end{cases}$$

We say that μ is **finite** if $\mu(\Omega) < \infty$, and it is **σ-finite** if Ω can be written as the union of countably many measurable sets of finite measure.

A set $E \in \mathcal{M}$ such that $\mu(E) = 0$ is called a **μ-null set**, or usually just a **null set** if the measure is clear from context. A property is said to hold **μ-almost everywhere** (often abbreviated a.e.) if the set of points for which it does not hold is μ-null (again, one would usually just write **almost everywhere** unless there was ambiguity).[1]

We now give some basic properties of measures.

Proposition 1. *If $E, F \in \mathcal{M}$ and $E \subseteq F$, then $\mu(E) \leq \mu(F)$.*

This property is called **monotonicity**.

Proof. Suppose $E, F \in \mathcal{M}$ and $E \subseteq F$. Then

$$\mu(F) = \mu(E \dot{\cup} (F \setminus E)) = \mu(E) + \mu(F \setminus E) \geq \mu(E)$$

as claimed. □

Note that monotonicity implies

$$0 = \mu(\varnothing) \leq \mu(E) \leq \mu(\Omega)$$

for every $E \in \mathcal{M}$, since $\varnothing \subseteq E \subseteq \Omega$.

Proposition 2. *For any countable collection of sets $\{E_i\} \subseteq \mathcal{M}$ (disjoint or not),*

$$\mu\left(\bigcup_i E_i\right) \leq \sum_i \mu(E_i)$$

This property is called **sub-additivity**.

Proof. Define $F_1 = E_1$ and $F_i = E_i \setminus (\bigcup_{j<i} E_j)$ for $i > 1$, noting that $\bigcup_{j \leq i} F_j = \bigcup_{j \leq i} E_j$ for all i and the F_i are disjoint. Then

$$\mu\left(\bigcup_i E_i\right) = \mu\left(\bigcup_i F_i\right) = \sum_i \mu(F_i) \leq \sum_i \mu(E_i)$$

where the last inequality follows by monotonicity since $F_i \subseteq E_i$ for all i. □

4 Lebesgue measure

Lebesgue measure is the measure that corresponds to our intuitive notion of physical size. For example, the Lebesgue measure of a measurable subset of $\mathbb{R}$ gives a number interpretable as the set's

[1] In order for this notion to be interesting we needed to first introduce a measure with nontrivial null sets, so we wait to give an example in the Lebesgue measure section.

length. Lebesgue measure can also be defined in higher dimensions (we omit this generalization), where it represents the area, volume, etc. of the set.

The strategy for constructing Lebesgue measure is to define it first on intervals, which have an obvious measure (their length), and then use that definition to define the measure of more complicated sets.

Let $\mathcal{I}$ be the set of all intervals (open, closed, or semi-open) on $\mathbb{R}$. Define $\ell : \mathcal{I} \to [0, \infty]$ by

$$\ell([a, b]) = b - a$$

with the same definition when $[a, b]$ is replaced by (a, b), $[a, b)$, or $(a, b]$. For infinite intervals, use the "obvious" convention that $\infty - a = \infty$ and $b - (-\infty) = \infty$.

The key tool in constructing Lebesgue measure is the **Lebesgue outer measure** $\lambda^* : \mathcal{P}(\mathbb{R}) \to [0, \infty]$, which is given by

$$\lambda^*(E) = \inf \left\{ \sum_{k=1}^{\infty} \ell(I_k) : I_k \in \mathcal{I}, E \subseteq \bigcup_{k=1}^{\infty} I_k \right\}$$

A set $E \subseteq \mathbb{R}$ is said to be **Lebesgue measurable** if for every $A \subseteq \mathbb{R}$,

$$\lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap E^c)$$

It turns out that the set of Lebesgue measurable sets is very large and contains pretty much any reasonable set that one would encounter in practice. However, it is possible[2] to construct pathological subsets of $\mathbb{R}$ that are not Lebesgue measurable.

The Lebesgue outer measure and Lebesgue measurable sets have a number of nice properties:

(i) The set of Lebesgue measurable sets, denoted $\mathcal{L}$, is a σ-algebra.

(ii) $\lambda^*|_{\mathcal{L}}$ is a measure on $\mathcal{L}$.

(iii) $\lambda^*|_{\mathcal{I}} = \ell$, so the measure agrees with our initial notion of interval length.

Defining the Lebesgue measure $\lambda = \lambda^*|_{\mathcal{L}}$, we have a measure space $(\mathbb{R}, \mathcal{L}, \lambda)$.[3]

4.0.1 Sets of measure zero

Consider the following intriguing property of Lebesgue measure.

Proposition 3. *If $E \subseteq \mathbb{R}$ is countable, then $\lambda(E) = 0$.*

Proof. First note that for any $x \in \mathbb{R}$, we have

$$\lambda(\{x\}) = \lambda([x, x]) = x - x = 0$$

Now suppose $E \subseteq \mathbb{R}$ is countable. Then we can write $E = \dot{\bigcup}_i \{x_i\}$, whence it follows that

$$\lambda(E) = \sum_i \lambda(\{x_i\}) = \sum_i 0 = 0$$

by the countable additivity of measures. □

[2] assuming the axiom of choice

[3] One can show that λ is the unique measure on $(\mathbb{R}, \mathcal{L})$ that extends ℓ. It turns out that uniqueness stems from the fact that ℓ is σ-finite.

Specifically, it may be surprising to consider that $\lambda(\mathbb{Q}) = 0$. It turns out that it's also possible to construct uncountable subsets of $\mathbb{R}$ that have Lebesgue measure zero, e.g. the Cantor set [**?**].

Now for the promised example of *almost everywhere*: the absolute value function $x \mapsto |x|$ is differentiable almost everywhere, since it is only not differentiable at $x = 0$, and $\lambda(\{0\}) = 0$. Note that Lebesgue measure is in some sense the "default" measure on $\mathbb{R}$, in that if no measure is specified (as in the previous sentence), the author is generally speaking in reference to Lebesgue measure.

5 Lebesgue integration

In this section we consider the problem of defining the integral of functions on an abstract measure space $(\Omega, \mathcal{M}, \mu)$.

Just as not all sets are measurable, not all functions are measurable. A function $f : \Omega \to \mathbb{R}$ is **measurable** if

$$\{\omega \in \Omega : f(\omega) \leq x\} \in \mathcal{M} \qquad \forall x \in \mathbb{R}$$

We follow the common approach of defining the Lebesgue integral for increasingly complicated functions in terms of integrals of simpler functions. The simplest functions to integrate are the **indicator functions**; if $E \in \mathcal{M}$, then its indicator function is

$$1_E(\omega) = \begin{cases} 1 & \omega \in E \\ 0 & \omega \notin E \end{cases}$$

The integral of an indicator function is defined as

$$\int_\Omega 1_E \, \mathrm{d}\mu = \mu(E)$$

From indicator functions we can construct **non-negative simple functions**, which are finite linear combinations of indicator functions:

$$\phi = \sum_{i=1}^{n} \alpha_i 1_{E_i}$$

where $\alpha_i \geq 0$ for all i. Here the integral is defined to be

$$\int_\Omega \phi \, \mathrm{d}\mu = \sum_{i=1}^{n} \alpha_i \int_\Omega 1_{E_i} \, \mathrm{d}\mu = \sum_{i=1}^{n} \alpha_i \mu(E_i)$$

and we use the convention that $0 \cdot \infty = 0$. Then we can define the integral of an arbitrary non-negative measurable function f as follows:

$$\int_\Omega f \, \mathrm{d}\mu = \sup \left\{ \int_\Omega \phi \, \mathrm{d}\mu : 0 \leq \phi \leq f, \phi \text{ simple} \right\}$$

Finally, we can extend the definition to arbitrary measurable functions by using the decomposition

$$f = f^+ - f^-$$

where

$$f^+(x) = \max(f(x), 0)$$
$$f^-(x) = \max(-f(x), 0)$$

If at least one of $\int_\Omega f^+ \, \mathrm{d}\mu$ and $\int_\Omega f^- \, \mathrm{d}\mu$ is finite, we define

$$\int_\Omega f \, \mathrm{d}\mu = \int_\Omega f^+ \, \mathrm{d}\mu - \int_\Omega f^- \, \mathrm{d}\mu$$

Furthermore if $\int_\Omega |f| \, \mathrm{d}\mu < \infty$, we say that f is **Lebesgue integrable**. Note that this is a slightly stronger condition than what is required for the previous definition; clearly $|f| = f^+ + f^-$, so f is Lebesgue integrable iff the integrals of both f^+ and f^- are finite.[4]

5.1 The Lebesgue integral on $\mathbb{R}$

We now consider the special case where $\Omega = \mathbb{R}$ and $\mu = \lambda$. In addition to being a very important special case of the general theory above, this scenario has a geometric interpretation that helps us better understand Lebesgue integration.

5.2 Comparison with the Riemann integral

In a nutshell, the Lebesgue integral is in many ways superior to the Riemann integral.

First, any function that is Riemann integrable on a bounded interval is also Lebesgue integrable, and the values of the integrals agree. But there exist functions are Lebesgue integrable but not Riemann integrable. For example, consider the rational indicator $1_\mathbb{Q}$ on $[0, 1]$. We know that for the Lebesgue integral,

$$\int_{[0,1]} 1_\mathbb{Q} \, \mathrm{d}\lambda = \lambda(\mathbb{Q} \cap [0, 1]) = 0$$

However it is easy to check that $1_\mathbb{Q}$ is not Riemann integrable: every non-trivial interval will contain at least one rational number and at least one irrational number, so no matter how the partition is chosen, the lower Darboux sum will be zero and the upper Darboux sum will be one.

But this is a rather contrived example. Of more practical importance is the existence of stronger convergence theorems, such as the monotone convergence theorem and dominated convergence theorem.

Another advantage of the Lebesgue integral, which admittedly is less important for our purposes, is that integration can be defined on spaces other than Euclidean space. The Riemann integral relies heavily on properties of the real line.

6 Probability

Suppose we have some sort of randomized experiment (e.g. a coin toss, die roll) that has a fixed set of possible outcomes. This set is called the **sample space** and denoted Ω.

We would like to define probabilities for some **events**, which are subsets of Ω. The set of events is denoted $\mathcal{F}$ and is required to be a σ-algebra.

Then we can define a **probability measure** $\pm : \mathcal{F} \to [0, 1]$ which must satisfy $\mathbb{P}(\Omega) = 1$ in addition to the axioms for general measures. The triple $(\Omega, \mathcal{F}, \pm)$ is called a **probability space**.[5]

[4] Why is the integral defined even for some functions that are not "integrable"? I'm not sure, and would love to know if anyone has more info. But all the sources I've consulted agree on these definitions.

[5] Note that a probability space is simply a measure space in which the measure of the whole space equals 1.

If $\mathbb{P}(A) = 1$, we say that A occurs **almost surely** (often abbreviated a.s.).[6] Conversely if $\mathbb{P}(A) = 0$, we say that A occurs **almost never**.

From these axioms, a number of useful rules can be derived.

Proposition 4. *If A is an event, then $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*

Proof. Using the countable additivity of $\pm$, we have

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \dot\cup A^c) = \mathbb{P}(\Omega) = 1$$

which proves the result. □

Proposition 5. *Let A be an event. Then*

(i) If B is an event and $B \subseteq A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

(ii) $0 = \mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$

Proof. (i) follows immediately from the monotonocity of measures. For (ii): the middle inequality follows from (i) since $\varnothing \subseteq A \subseteq \Omega$. We also have $\mathbb{P}(\varnothing) = 0$ by applying the previous proposition with $A = \Omega$. □

Proposition 6. *If A and B are events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*

Proof. The key is to break the events up into their various overlapping and non-overlapping parts.

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B) \dot\cup (A \setminus B) \dot\cup (B \setminus A)) \\
&= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) \\
&= \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)
\end{aligned}$$

□

Proposition 7. *If $\{A_i\} \subseteq \mathcal{F}$ is a countable set of events, disjoint or not, then*

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i)$$

This inequality is sometimes referred to as **Boole's inequality** or the **union bound**.

Proof. Follows immediately from the sub-additivity of measures. □

6.1 Random variables

Intuitively, a **random variable** is some uncertain quantity with an associated probability distribution over the values it can assume.

Formally, a random variable on a probability space $(\Omega, \mathcal{F}, \pm)$ is a measurable function $X : \Omega \to \mathbb{R}$.[7]

[6] This is a probabilist's version of the measure-theoretic term *almost everywhere*.

[7] More generally, the codomain can be any measurable space, but $\mathbb{R}$ is the most common case by far and sufficient for our purposes.

We denote the range of X by $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. To give a concrete example (taken from [**?**]), suppose X is the number of heads in two tosses of a fair coin. The sample space is

$$\Omega = \{hh, tt, ht, th\}$$

and X is determined completely by the outcome ω, i.e. $X = X(\omega)$. For example, the event $X = 1$ is the set of outcomes $\{ht, th\}$.

It is common to talk about the values of a random variable without directly referencing its sample space. The two are related by the following definition: the event that the value of X lies in some set $S \subseteq \mathbb{R}$ is

$$X \in S = X^{-1}(S) = \{\omega \in \Omega : X(\omega) \in S\}$$

Here the X^{-1} notation means the preimage of S under X, not the inverse of X.

Note that special cases of this definition include X being equal to, less than, or greater than some specified value. For example

$$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(\{x\})) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

6.1.1 The cumulative distribution function

The **cumulative distribution function** (c.d.f.) gives the probability that a random variable is at most a certain value:

$$F(x) = \mathbb{P}(X \leq x)$$

The c.d.f. can be used to give the probability that a variable lies within a certain range:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a)$$

6.1.2 Discrete random variables

A **discrete random variable** is a random variable that has a countable range and assumes each value in this range with positive probability. Discrete random variables are completely specified by their **probability mass function** (p.m.f.) $p : X(\Omega) \to [0, 1]$ which satisfies

$$\sum_x p(x) = 1$$

For a discrete X, the probability of a particular value is given exactly by its p.m.f.:

$$\mathbb{P}(X = x) = p(x)$$

In fact, any nonnegative function that sums to one over a countable domain induces a discrete probability space.

Proposition 8. *Suppose Ω is a non-empty countable set and $p : \Omega \to [0, 1]$ is such that $\sum_{\omega \in \Omega} p(\omega) = 1$. Let $\mathcal{F} = \mathcal{P}(\Omega)$ and*

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$$

for any event $A \in \mathcal{F}$. Then

(i) $(\Omega, \mathcal{F}, \pm)$ is a probability space.

(ii) If $S \subset \mathbb{R}$ with $|S| = |\Omega|$, then any bijection $X : \Omega \to S$ is a random variable on this space with probability mass function $p \circ X^{-1}$.

Proof. $\mathcal{F}$ is clearly a σ-algebra since it contains every subset of Ω and thus is closed under all complements and unions. Thus all that must be shown is that $\pm$ is a probability measure. We have $\mathbb{P}(\Omega) = \sum_{\omega \in \Omega} p(\omega) = 1$ immediately by assumption. To show countable additivity, we see that if $\{A_i\} \subseteq \mathcal{F}$ are disjoint, then

$$\mathbb{P}\left(\dot{\bigcup_i} A_i\right) = \sum_{\omega \in \dot{\bigcup}_i A_i} p(\omega) = \sum_i \sum_{\omega \in A_i} p(\omega) = \sum_i \mathbb{P}(A_i)$$

which proves (i).

To show (ii), suppose $S \subset \mathbb{R}$ with $|S| = |\Omega|$ and let $X : \Omega \to S$ be a bijection. It is clear that X is measurable, again because $\mathcal{F}$ contains every subset of Ω. We also have for any $x \in S$,

$$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(\{x\})) = \mathbb{P}(\{X^{-1}(x)\}) = p(X^{-1}(x)) = (p \circ X^{-1})(x)$$

so $p \circ X^{-1}$ is the probability mass function of X. □

6.1.3 Continuous random variables

A **continuous random variable** is a random variable that has an uncountable range and assumes each value in this range with probability zero. Most of the continuous random variables that one would encounter in practice are **absolutely continuous random variables**[8], which means that there exists a function $p : \mathbb{R} \to [0, \infty)$ that satisfies

$$F(x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}z$$

The function p is called a **probability density function** (abbreviated p.d.f.) and must satisfy

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

The values of this function are not themselves probabilities, since they could exceed 1. However, they do have a couple of reasonable interpretations. One is as relative probabilities; even though the probability of each particular value being picked is technically zero, some points are still in a sense more likely than others.

One can also think of the density as determining the probability that the variable will lie in a small range about a given value. Recall that for small ϵ,

$$\mathbb{P}(x - \epsilon/2 \leq X \leq x + \epsilon/2) = \int_{x-\epsilon/2}^{x+\epsilon/2} p(z)\,\mathrm{d}z \approx \epsilon p(x)$$

using a midpoint approximation to the integral.

Here are some useful identities that follow from the definitions above:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\,\mathrm{d}x$$
$$p(x) = F'(x)$$

[8] Random variables that are continuous but not absolutely continuous are called **singular random variables**. We will not discuss them, assuming rather that all continuous random variables admit a density function.

6.1.4 Other kinds of random variables

There are random variables that are neither discrete nor continuous. For example, consider a random variable determined as follows: flip a fair coin, then the value is zero if it comes up heads, otherwise draw a number uniformly at random from $[1, 2]$. Such a random variable can take on uncountably many values, but only finitely many of these with positive probability. We will not discuss such random variables.

References

[1] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications (Second Edition)*. New York: John Wiley & Sons, 1999.

[2] J. Pitman, *Probability*. New York: Springer-Verlag, 1993.

[3] J. S. Rosenthal, *A First Look at Rigorous Probability Theory (Second Edition)*. Singapore: World Scientific Publishing, 2006.