A Measure Theory Tutorial (*Measure Theory for Dummies*)

Maya R. Gupta
`{gupta}@ee.washington.edu`

Dept of EE, University of Washington
Seattle WA, 98195-2500

UW
Electrical
Engineering

UWEE Technical Report
Number UWEETR-2006-0008
May 2006

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

A Measure Theory Tutorial (*Measure Theory for Dummies*)

Maya R. Gupta
{gupta}@ee.washington.edu

Dept of EE, University of Washington
Seattle WA, 98195-2500

University of Washington, Dept. of EE, UWEETR-2006-0008
May 2006

Abstract

This tutorial is an informal introduction to measure theory for people who are interested in reading papers that use measure theory. The tutorial assumes one has had at least a year of college-level calculus, some graduate level exposure to random processes, and familiarity with terms like "closed" and "open." The focus is on the terms and ideas relevant to applied probability and information theory. There are no proofs and no exercises.

Measure theory is a bit like grammar, many people communicate clearly without worrying about all the details, but the details do exist and for good reasons. There are a number of great texts that do measure theory justice. This is not one of them. Rather this is a hack way to get the basic ideas down so you can read through research papers and follow what's going on. Hopefully, you'll get curious and excited enough about the details to check out some of the references for a deeper understanding.

A Something to measure

First, we need something to measure. So we define a "measurable space." A measurable space is a collection of events $\mathcal{B}$, and the set of all those events Ω, which is called the event space or sample space. Why do you need Ω? For one, having an event space makes it possible to define complements of sets; if $F \in \Omega$, then when we say F^C we mean the set of events in Ω that are disjoint from F. A measurable space is written $(\Omega, \mathcal{B})$.

A.1 Algebras and Fields

Often, you will see that the collection of events $\mathcal{B}$ in a measurable space is a σ-algebra. A σ-algebra is a special kind of collection of subsets of the sample space Ω: a σ-algebra is complete in that if some set A is in your σ-algebra, then you have to have A^C (the complement of A) in your set too. Also, it must be that if you have two sets A and B in your collection of sets, then the union $A \cup B$ must also be in your collection of sets. There are other special types of collections of sets that you may run into, for example, a $\sigma - field$. The smallest possible σ-field is a collection of just two sets, $\{\Omega, \emptyset\}$. The largest possible σ-field is the collection of all the possible subsets of Ω, this is called the powerset.

B Measure

A measure μ takes a set A (from a measurable collection of sets $\mathcal{B}$), and returns "the measure of A," which is some positive real number. So ones writes $\mu : \mathcal{B} \rightarrow [0, \infty)$. An example measure is volume, which goes by the name Lebesgue measure. In general, measures are generalized notions of volume. The triple $(\Omega, \mathcal{B}, \mu)$ combines a measurable space and a measure, and thus the triple is called a *measure space*. A measure is defined by two properties:

1. Nonnegativity: $\mu(A) \geq 0$ for all $A \in \mathcal{B}$

2. Countable Additivity: If $A_i \in \mathcal{B}$ are disjoint sets for $i = 1, 2, \ldots$, then the measure of the union of the A_i is equal to the sum of the measures of the A_i.

You can see how our ordinary notion of volume satisfies these two properties. There are a couple variations on measure that you will run into. One is a *signed measure*, which can be negative. A special case of measure is the probability measure. A probability space is just a measure space with a probability measure. And a probability measure P has the two above properties of a measure but it's also normalized, such that $P(\Omega) = 1$.

A probability measure P over discrete set of events is basically what you know as a probability mass function. For example given probability measure P and two sets $A, B \in \mathcal{B}$, we can familiarly write

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

B.1 Measure zero

A set of measure zero is some set $A \in \mathcal{B}$ such that $\mu(A) = 0$. In terms of probability, that means "I've got this event A in my sample space, but it's never going to happen." The nice thing about sets of measure zero is they don't count when you want to state things about the collection of sets $\mathcal{B}$. For example, you could integrate a function f over $\mathcal{B}$, and the value $f(A)$ might be infinite, but you can ignore $f(A)$ if set A has measure zero.

B.2 Measurable functions

A function defined over a measurable set is called a *measurable function*.

B.3 Borel sets

A σ-algebra (collection of sets) that appears often is the Borel σ-algebra. You'll usually see people talk about "the Borel σ-algebra on the real line," which is the collection of sets that is the smallest sigma-algebra that includes the open subsets of the real line. A Borel set is an element of a Borel σ-algebra. It turns out that just about any set you can describe on the real line is a Borel set, for example, the unit line segment $[0, 1]$ is a Borel set, the irrational numbers form a Borel set, etc.

You can define the sample space to be the real line $\mathbb{R}$, and then the corresponding Borel σ-algebra is the collection of sets $\mathcal{R}$. Thus, $(\mathbb{R}, \mathcal{R})$ is a measurable space. You could also define a Borel measurable space for $\mathbb{R}^2$, etc., as $(\mathbb{R}^d, \mathcal{R}^d)$. One thing that makes the Borel sets so powerful is that if you know what a probability measure does on every interval, then you know what it does on all the Borel sets. This shows up in proofs where people want to say that some measure μ_1 is really equivalent to some other measure μ_2. To do that they just need to show that μ_1 and μ_2 are equivalent on all intervals, and then they have proven that the two measures are equivalent for all the Borel sets (and hence over the measurable space). The Borel set can be equivalently defined by intervals of various types, for example, you could use the set of all open intervals $(a, b), -\infty \leq a \leq b \leq \infty$, or all closed intervals of the real line $[a, b], -\infty \leq a \leq b \leq \infty$. Or, you could use all right closed intervals: $(-\infty, b], -\infty \leq b \leq \infty$. In fact, that's how one defines a cummulative distribution function, and this is the basis for the math that tells you that if two cdfs are equal for all choices of b, then the two probability measures must be equal.

B.4 Support

The support of a measure is all the sets that do not have measure zero. For example you might say, "the probability measure μ has support only on the unit interval," by which you mean there is zero probability of drawing a point bigger than one or smaller than zero. You often see written "the measure has compact support" to note that the support of the measure forms a compact (=closed and bounded) set.

B.5 Lebesgue measure

The Lebesgue measure $\mu_L(A)$ is just the volume (or area, or length) of set A. For example, $\mu_L([0, 1]) = 1$.

B.6 Borel measure

To call a measure a Borel measure means it is defined over a Borel σ-algebra.

B.7 Example

This example is based on an example from Capinski and Kopp's book, page 45 [6] of what it means to say "draw a number from $[0,1]$ at random."
Restrict Lebesgue measure m to the interval $B = [0,1]$ and consider the σ-field $\mathcal{M}$ of measurable subsets of $[0,1]$. Then $m_{[0,1]}$ is a probability measure on $\mathcal{M}$. Since all subintervals of $[0,1]$ with the same length have the same measure, the mass of $m_{[0,1]}$ is spread uniformly over $[0,1]$, so that the measure of $[0,1/10)$ is the same as the measure of $[6/10,7/10)$ (both are $1/10$). Thus all numerals are equally likely to appear as first digits of the decimal expansion of a number drawn randomly according to this measure.

C The truth about *Random Variables*

A basic definition of a random variable is that it specifies a set of events that happen with corresponding probabilities. More formally, let there be some measure space $(\Omega, \mathcal{F}, P)$, then a random variable is a measurable function X that maps the measurable space $(\Omega, \mathcal{F})$ to another measurable space, usually the Borel σ-algebra of the real numbers $(\mathbb{R}, \mathcal{R})$.

We can see that the formal definition is saying the same thing as the basic definition. Consider an output event $A \in \mathcal{R}$. The random variable X induces a probability on the output event A, which is $\mathbb{P}(A) = P(\{w : X(w) \in A\})$. That is, there is some set of events $\{w\}$ that X maps to the output event A, and the probability of output event A is the total probability of those events $\{w\}$. That is, the probability of a Borel set A in the output space is equal to the probability of the inverse image under X of that Borel set:

$$\mathbb{P}(A) = P(X^{-1}(A)) = P(\{w : X(w) \in A\}).$$

This works for every Borel set in the output space, so the random variable X induces a probability measure over the space.

As shorthand, one writes the probability $\mathbb{P}(A) = P(X \in A)$. In fact, it's common not to write the induced measure at all, and just write $P(X \in A)$.

C.1 Distributions

The probability measure $\mathbb{P}$ over the output measurable space induced by a random variable X is called the *distribution* of X [7]. However, the term *distribution* is also used in a more specific way. As we foreshadowed in the section on Borel sets, the complete description of the probability measure induced by a random variable X requires knowledge about $P(X \in A)$ for all sets $A \in \mathcal{R}$. However, since the Borel σ-algebra can be generated by the set of intervals $(-\infty, x)$ for all $x \in \mathcal{R}$, we only have to know $P(X \in A)$ for every set $A = (-\infty, x)$. Then, the *distribution function* of X is $F(x) = P(X \leq x)$. The distribution function is usually indexed by the random variable, such as F_X or F_Y.

Then one can say that the induced probability measure over the interval $(a,b]$ is $\mathbb{P}((a,b]) = F(b) - F(a)$.

C.2 Probability densities

Given a distribution, there might not be a corresponding density. To have a density you need the distribution function F to be what is called "absolutely continuous," that is, it must be that for all $a < b$,

$$F(b) - F(a) = \int_a^b f(x)dx,$$

where f is a non-negative, Lebesgue measurable function. If f exists such that the above holds, then f is called the density of the distribution F.

C.3 Discrete distributions

A discrete distribution F has the familiar corresponding point mass function, or probability mass function. For a discrete distribution F there is some countable set of numbers $\{x_j\}$ and point masses $\{p_j\}$ such that

$$F(x) = \sum_{x_j \leq x} p_j,$$

for all $x \in \mathcal{R}$. The $\{p_j\}$ form the probability mass function (pmf) over the events, which are defined to be intervals of the real line.

C.4 Expectation and integration

Expectation and integration are fundamentally the same thing. We are going to start with expectation (the average) as the more fundamental concept, and from there develop integration.

Here is a limit definition of expectation for ***non-negative*** random variables X [5][7]. This is our starting definition, from which we will get to the familiar integral formula:

$$EX = \lim_{n \to \infty} \sum_{k=1}^{n2^n} \frac{k-1}{2^n} P\left(\frac{k-1}{2^n} \leq X < \frac{k}{2^n} \right),$$

(note this limit might not exist).

This formula has all the parts you expect: it's a sum over an increasingly larger number of increasingly smaller components, and for each component one takes the measure of the interval. What about negative random variables? We'll need a couple extra definitions. Let $X^+ = max(X, 0)$, and let $X^- = -min(X, 0)$. Then, for an arbitrary random variable X let $EX = EX^+ - EX^-$ as long as EX^+ and/or EX^- are finite. Note that EX^- turns a negative X into a positive, but we then negafy it, so the negative aspect is not lost.

This general definition of expectation is the subtraction of two limit-sums, and we give this general definition of expectation a new notation, the integral notation:

$$EX = \int_{\Omega} X(w) dP(w)$$

which is also written

$$EX = \int X dP,$$

and sometimes written

$$EX = \int X(x) P(dx).$$

If $E|X| < \infty$, then one says that the random variable X is integrable. You'll note from the limit-sum definition that if one takes the integral of a set of measure 0, one gets 0. That is a simple but key idea in many proofs, and you'll often see equivalence relationships where some q *equals* some p if they agree on all sets that do not have measure 0.

The most common measure to use in integration is the Lebesgue measure, which is for almost all practical purposes equivalent to the standard Riemann integration that one first learns. As noted in Section E, Riemann integration has some problems that make it not as useful as Lebesgue integration, and the reader is referred (for example) to Capinski and Kopp for more details [6].

D Entropy

Entropy is a useful function of a random variable and in this section we will use it to solidify some of the ideas and notation introduced above. First, consider the entropy of a discrete alphabet random variable f defined on the probability space $(\Omega, \mathcal{B}, P)$. Then the entropy is

$$H_P(f) = -\sum_{a \in A} P(f = a) \ln P(f = a).$$

Also, f induces a probability mass function (pmf) p_f, where $p_f = P(w : f(w) = a) = P(f = a)$, so you can equivalently write

$$H_P(f) = -\sum_{a \in A} p_f(a) \ln p_f(a).$$

A discrete random variable f induces a partition on the input space Ω that corresponds to the inverse image of each event: let the partition $\mathcal{Q}$ consist of sets $\{Q_i : i = 1, 2, \ldots, \|A\|\}$ where $Q_i = \{w : f(w) = a_i\} = f^{-1}(\{a_i\})$. Then you can also write entropy in terms of the induced partition $\mathcal{Q}$:

$$H_P(\mathcal{Q}) = -\sum_{i=1}^{\|A\|} P(Q_i) \ln P(Q_i).$$

E Limits

To quote Gut [5], "One of the basic questions in mathematics is to what extent limits of objects carry over to limits of functions of objects." One of the more important results in this area is Lebesgue's Dominated Convergence Theorem. A formal statement of this can be found on mathworld's page (www.mathworld.com). Basically, it says that the integral of a function f with a measure μ is the same as the limit of the integral of f_n, where $f_1, f_2, \ldots, f_n$ is a sequence of measurable functions that converges to f. There are some other restrictions on the f_n's (see the formal statement). What's powerful about this theorem though is that one doesn't have to assume that f is measurable, instead, the theorem concludes that f is integrable, and shows you how to integrate it by instead taking the limit of the integral of the sequence of functions. The integration one learns as a kid is Riemann integration. This theorem doesn't work for Riemann integration, and that is considered one of the flaws of Riemann integration that makes Lebesgue integration more general and more useful. For more on the flaws of Riemann integration, see for example [6].

F Read on

If you are interested in a more thorough understanding of measure theory and probability, one of the friendliest books is Resnick's [8], which teaches measure theoretic graduate level probability with the assumption that you do not have a B.A. in mathematics. Other good texts are an undergraduate text on measure theory [6], and Gut's graduate level measure-theoretic probability book [5]. There are certainly plenty of other probability and measure theory text books, but these three are relatively well-suited for self-study. If you've decided you aren't so interested in formal probability, but want to learn more about approaches to solving probability problems, I recommend Richard Hamming's book [2].

If you are interested in information theory you can solidify your understanding of the use of measure theory in information theory by reading Bob Gray's book [7], Kullback's book [1], or the information theory book by Ash [4] (you might want to read the less formal book by Reza [3] before Ash). Gray's book is available free on-line, and the Kullback, Ash, and Reza books are available in inexpensive Dover editions.

References

[1] S. Kullback, "Information theory and statistics", *Dover*, 1997.

[2] R. W. Hamming, "The art of probability for scientists and engineers", *Addison Wesley*, 1993.

[3] F. M. Reza, "An introduction to information theory", *Dover*, 1994.

[4] R. B. Ash, "Information theory", *Dover*, 1990.

[5] A. Gut, "Probability: A Graduate Course", *Springer*, 2005.

[6] M. Capinski and E. Kopp, "Measure, Integral, and Probability", *Springer Undergraduate Mathematics Series*, 2004.

[7] R. M. Gray, "Entropy and Information Theory", *Springer Verlag (available free online)*, 1990.

[8] S. I. Resnick, “A probability path”, *Birkhäuser*, 1999.

[9] T. M. Cover and J. A. Thomas, “Elements of Information Theory”, *Wiley Series in Telecommunications*, 1991.