The Lebesgue Integral, Chebyshev's Inequality, and the Weierstrass Approximation Theorem

George Stepaniants

June 6, 2017

Contents

Abstract

In this paper, we will prove a famous theorem known as the Weierstrass Approximation Theorem. In 1885, Weierstrass (being 70 years of age) proved a rather astounding theorem that on the interval $[0, 1]$, any continuous function can be approximated infinitely close by a polynomial function. His proof heavily relied on intricate analysis concepts and involved building up a sequence of polynomials from a convolution with a Gaussian heat kernel. The proof we display in this paper, is not of Weierstrass, but of Bernstein. In 1912, Sergei Bernstein introduced his Bernstein polynomials to prove this theorem and used an elegant probabilistic argument. His argument involved the use of Chebyshev's Inequality which we will shall also prove in this paper. Our rendition of Bernstein's proof is taken from Kenneth Levasseur's short paper in The American Mathematical Monthly [3].

In order to prove Chebyshev's Inequality, we will introduce some measure theory in order to define Lebesgue measure and Lebesgue integration. Some of our measure-theoretic definitions involving σ-algebras and meaure spaces were taken from the Camridge University class notes of "Probability and Measure" by J. R. Norris [4]. Our definition of Lebesgue integration will follow the Daniell-Riesz approach that is described in the "Lebesgue Integral for Undergraduates" text written by W. Johnston [2]. This approach does not attempt to introduce the reader to complicated measure theoretic concepts but instead defines the Lebesgue integral by defining what it means to integrate over step functions and approximating any function below by a sequence of nondecreasing step functions. After defining the Lebesgue integral, we will move on to define basic probabilistic concepts such as a probability space, probability measure, random variables, density functions, distributions, expected value, and variance. Some of the examples involving Lebesgue measure and probability are taken from Botts' paper on "Probability Theory and the Lebesgue Integral" [1]. We will then introduce two probability distributions and we will

compute the mean and variance of the binomial distribution. Finally, we will prove Chebyshev's Inequality in its most general form and will apply it in Bernstein's proof of the Weierstrass Approximation Theorem.

1 Introduction of Concepts

Here we give a broad overview of the topics presented in our paper and how they build to prove the Weierstrass Approximation Theorem.

To begin, Section 2 of this paper introduces basic measure theoretic concepts. It first gives the definition of a power set and uses this to define a σ-algebra which is essentially a subset of a power set. Every set in the σ-algebra is defined to be a *measurable set* which means that there exists some way to assign a real positive number from 0 to ∞ to every such set. There exist many ways to assign real numbers (essentially lengths or volumes) to sets in the σ-algebra and a function that does this is called a *measure*. A *measurable space* is a set paired with its σ-algebra. We continue by defining a *measure space* which is effectively a *measurable space* equipped with a particular choice of measure. Lastly, we state the definition of a *measurable map* which is a function that maps a measurable space to a measurable space. After this, we introduce a particular type of measure called the *Lebesgue measure* which we define on the reals. We use the Daniell-Riesz approach [2] to introduce *Lebesgue integration* and by defining Lebesgue integration of step functions and approximating any function from below by a nondecreasing sequence of step functions. This allows us, in limit, to define the integral for any function that can be approximated by a sequence of step functions from below. The space of such functions is called L^0. We then go on to show that the space L^0 is not linear (not closed under linear operations) and show that the L^0 space is not a good space to define all Lebesgue integrable functions. Therefore, we create a new space of functions called L^1 which turns out to be equivalent to the space of Lebesgue integrable functions although we do not prove this result in the paper.

After introducing the results above, Section 3 focuses on probabilistic theory. We give the definition of a *probability space* which is a measure space where the measure defined on it is a *probability measure*. A probability measure is essentially a measure that assigns every set a real number from 0 to 1 and gives the entire sample space (universal set) a measure of 1. We then go on to define a *random variable* which is essentially a measurable map with an additional restriction imposed on it. Random variables can be either discrete or continuous. We first examine these cases separately and later state that discrete and continuous cases of the same concept (random variables, distributions) in probability theory can be unified using the Dirac delta function. We introduce the concept of a *probability mass function* which assigns a probability to every discrete value of a random variable. We also define the *probability density function* in the continuous case which assigns an infinitesimal density to every value of a continuous random variable. Both the probability mass function and the probability density function allow us to compute the probability that a random variable hits certain values or is within a certain range through summation or Lebesgue integration respectively. Lastly, Section 3 concludes by defining what a *probability distribution* is and *mean* and *variance* of a probability distribution. Finally, we prove Chebyshev's Inequality in its most general measure theoretic representation and show how the probabilistic statement of Chebyshev's Inequality is a special case of this.

Finally, we prove the Weierstrass Approximation Theorem in Section 4 through a constructive proof using the Bernstein polynomials that were used in Bernstein's original proof [3] along with Chebyshev's Inequality.

2 Measure Theory and the Lebesgue Integral

2.1 Basics of Measure Theory

Definition 2.1 (Power Set)**.** Let X be some set. The power set of X, denoted as 2^X, is the set of all subsets of X.

Definition 2.2 (σ-algebra)**.** Let X be a set and let 2^X be its power set. Then a subset $\sigma \subseteq 2^X$ is called a σ-algebra for X if it satisfies the following three properties:

- X is in σ: $X \in \sigma$
- σ is closed under complementation: $A \in \sigma \Rightarrow X \setminus A \in \sigma$
- σ is closed under countable unions:
  $A_1, A_2, A_3, ... \in \sigma$ and $A_i \cap A_j = \emptyset, i \neq j$
  $\Rightarrow \bigcup_{i=1}^{\infty} A_i \in \sigma$

The σ-algebra for (or generated by) X can be denoted as $\sigma(X)$. One very simple (but not very useful) σ-algebra is the power set of X, namely 2^X. We will see in the subsequent section why σ-algebras are important in the context of probability. Effectively, we may may view our set X as a set of outcomes and $\sigma(X)$ would be the set of events where each event is in turn, a subset of outcomes of X.

Definition 2.3 (Measurable Set)**.** Let X be some set. If $A \in \sigma(X)$ then A is called a *measurable set* in X.

Definition 2.4 (Measurable Space)**.** If X is some set and σ is the σ-algebra generated by X, then the pair (X, σ) is called a *measurable space*.

In the next definitions, let $\bar{\mathbb{R}}$ denote the real numbers $\mathbb{R}$ with $+\infty$.

Definition 2.5 (Measure)**.** On the measurable space (X, σ), the map $\mu : X \to \bar{\mathbb{R}}$ is called a *measure* if it satisfies the following three properties:

- Null empty set: $\mu(\emptyset) = 0$
- Non-negativity: $\forall E \in \sigma, \mu(E) > 0$
- Countable/Sigma additivity: $A_1, A_2, A_3, ... \in \sigma$ and $A_i \cap A_j = \emptyset, i \neq j$
  $\Rightarrow \mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

There are many different measures that are used for very different purposes. The counting measure is $\mu(S) = |S|$ (number of elements in S) and is used in application with discrete probability distributions. It is mostly used on countable sets. The Lebsegue measure is defined to have the property that $\mu([0, 1]) = 1$. We will go into more detail on the Lebesgue measure and Lebesgue integration in the next section. Another type of measure that we will closely examine is the probability measure which takes the value 1 on the whole space and takes all its values in the unit interval $[0, 1]$.

Definition 2.6 (Measure Space)**.** Let X be a set, commonly referred to as the universal set. Let σ be a σ-algebra for X and μ be the measure on the measurable space (X, σ). The triple (X, σ, μ) is called a *measure space*.

Essentially, a measure space is just a measurable space equipped with a measure μ. The main example of a measure space that we will deal with is the probability space which is a measurable space equipped with the probability measure mentioned earlier.

Definition 2.7 (Measurable Map)**.** Let (X, σ_X) and (Y, σ_Y) be two measurable spaces. A map or function $f : X \to Y$ is called a *measurable map* if for every Y-measurable set $A \in \sigma_Y$, the inverse image is X measurable $f^{-}1(A) \in \sigma_X$.

2.2 Lebesgue Measure and the Daniell-Riesz Approach to Lebesgue Integration

Now we will provide a short and intuitive description of the Lebesgue measure. Instead of focusing on rigorously defining the Lebesgue measure, we will focus on noting its most important properties. We will then go on to describe the theory of Lebesgue integration using the Daniell-Riesz approach

[2] which bypasses many complicated measure-theoretic concepts to define Lebesgue integration through the use of step functions.

For consistency, we will refer to the Lebesgue measure in this section and all later sections with the symbol μ.

Definition 2.8 (Lebesgue Measure of an Interval)**.** The Lebesgue measure has the important property that the measure of an interval is its length regardless if the interval is open, closed, or half-open. Therefore,

$$\mu([a,b]) = \mu([a,b)) = \mu((a,b]) = \mu((a,b)) = b - a \tag{1}$$

Like any measure, the Lebesgue measure has the three properties of having a null empty set, is non-negative, and is countably additive.

Definition 2.9 (Lebesgue Measure Zero)**.** A set S has Lebesgue measure zero if it can be covered with a sequence of open intervals $I_1, I_2, I_3, ...$ so that the sum of the measures of all of the intervals (bounded total measure) can be made arbitrarily small. By the countable additivity property of a measure, we know that the bounded total measure is, in fact, the measure of $\bigcup_{n=1}^{\infty} I_n$. More formally,

$$\forall \epsilon > 0, \exists \{I_n\}_{n=1}^{\infty} \quad \sum_{n=1}^{\infty} \mu(I_n) < \epsilon \tag{2}$$

If a set S has Lebesgue measure zero, we write that $\mu(S) = 0$. It is relatively easy to see from this that any set with a finite number of real numbers has Lebesgue measure zero. A slightly less intuitive concept is that every countably infinite set of reals has Lebesgue measure zero [2]. By countably infinite, we mean that there is some way to enumerate this inifinite set of real numbers.

To see why this is true, let's assume that the set S is a countably infinite set of real numbers $a_1, a_2, a_3, ...$ and let $\epsilon > 0$ be arbitrary.

Define our sequence of open sets to be

$$I_n = (a_n - \frac{\epsilon}{2^{n+1}}, a_n + \frac{\epsilon}{2^{n+1}}) \tag{3}$$

Note that each a_n is at the center of the open set I_n so the union $\bigcup_{n=1}^{\infty} I_n$ covers our entire set of real numbers. Furthermore, the bounded total measure of our covering becomes

$$\sum_{n=1}^{\infty} \mu(I_n) = \sum_{n=1}^{\infty} ((a_n + \frac{\epsilon}{2^{n+1}}) - (a_n - \frac{\epsilon}{2^{n+1}})) = \sum_{n=1}^{\infty} \frac{\epsilon}{2^n} = \epsilon \frac{1/2}{1 - 1/2} = \epsilon \tag{4}$$

The last equality in the derivation is due to the sum of the geometric series which converges in our case since $\frac{1}{2} < 1$. Therefore, by definition, $\mu(S) = 0$.

Now we will introduce Lebesgue integration. Before we can do this, we must first familiarize ourselves with the step function and how we will use sums of step functions to approximate a class of functions know as *Lebesgue integrable functions*. The Daniell-Riesz approach first builds up a space of functions know as the L^0 space. As it turns out, the L^0 space is not linear. In other words, if two functions are in L^0, then by addition and scalar multiplication, we can find a linear combination of these two function that is not in L^0. Because of this, the Daniell-Riesz approach then goes on to create the L^1 space from the L^0 space and this new space is in fact, defined as the space of all Lebesgue integrable functions. Although we will not prove this result here, it can be shown that the space of Lebesgue integrable functions obtained through measure-theoretic definitions of the Lebesgue integral is equivalent to the space L^1 in the Daniell-Riesz approach.

Definition 2.10 (Characteristic Function)**.** A characteristic function on a set S is

$$\chi_S(x) = \begin{cases} 1, x \in S \\ 0, x \notin S \end{cases} \tag{5}$$

Definition 2.11 (Lebesgue Integral of Characteristic Function)**.** Let I be a bounded interval and c be a real constant. The Lebesgue integral of $c\chi_I(x)$ over $\mathbb{R} = (-\infty, \infty)$ is defined as

$$\int_{\mathbb{R}} c \cdot \chi_I(x) d\mu(x) = c \cdot \mu(I) \tag{6}$$

Here the $d\mu(x)$ in the integral denotes that our "dummy variable" of integration is x (which commonly denotes the real number line) and that we are integrating with respect to the Lebesgue measure μ. In the appropriate contexts, we will use either $d\mu(x)$, $d\mu$, or dx. In this section, we are defining Lebesgue integration on the reals so we will solely use dx in our notation.

Definition 2.12 (Step Function)**.** A step function f has the form

$$f(x) = \sum_{j=1}^{n} c_j \cdot \chi_{I_j}(x) \tag{7}$$

where c_j is a real constant corresponding to the bounded interval I_j.

Definition 2.13 (Lebesgue Integral of Step Function)**.** The Lebesgue integral of the step function $f(x) = \sum_{j=1}^{n} c_j \cdot \chi_{I_j}(x)$ is defined to be

$$\int_{\mathbb{R}} f(x) dx = \sum_{j=1}^{n} c_j \cdot \mu(I_j) \tag{8}$$

Now we will prove that any step function can be written as a sum of weighted characteristic functions defined on disjoint intervals [2].

Lemma 2.1. A step function $f = \sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x)$ with intervals T_j can be rewritten as $f = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$ where the intervals I_j are pairwise disjoint. In other words, $A_s \cap A_t = \emptyset, s \neq t$.

Proof. Since we have a finite number of intervals J_i, lay out the distinct endpoints of all the intervals in order. So, we can assume that the intervals J_i have r distinct endpoints on the real line which we will denote in order from least to greatest by $p_i, i = 1, 2, 3, ...$

Now, construct $m = 2r - 1$ intervals I_j. Let r of these intervals be trivial and of the form $[p_j, p_j]$ which only contain the single point $p_j, 1 \leq j \leq r$. Let the other $r - 1$ of these intervals be open and take on the form (p_j, p_{j+1}) where the intervals lie in between the adjacent endpoints. By construction, the intervals I_j we have created are disjoint.

For any of the intervals I_j, we know that either $I_j \subseteq J_i$ or $I_j \cap J_i = \emptyset$. Now, let $c_j = \sum k_i$ for all i where $I_j \subseteq J_i$. If $I_j \cap J_i \neq \emptyset$ for all i, then define $c_j = 0$.

Now, we want to prove that we can write our step function f equivalently as $f = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$. Take any real value x between the two outermost endpoints p_1 and p_r. This value x must lie in one of the intervals I_s. Therefore, since the intervals I_j are disjoint, we have that $\sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x) = c_s$. And we know by the way we constructed the coefficients of our step function that $c_s = \sum k_i$ for all i where $I_s \subseteq J_i$.

Thus, for any value $x \in [p_1, p_r]$ we have that if $x \in I_s$, then $f(x) = \sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x) = \sum k_i$ where $I_s \subseteq J_i$. Finally, $f(x) = c_s = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$. So,

$$f = \sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x) = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x) \tag{9}$$

where the I_j's are disjoint.

Note how in this proof we made our disjoint interval step function construction using intervals containing only one point and open intervals. We could simplify this step function by combining terms in our summation with immediately adjacent intervals that have the same constant coefficient. Our intervals would still be disjoint but our step function would now be constructed out of the fewest number of intervals. This minimum interval representation is, in fact, unique. □

To see an example of the results proved in the lemma above, let our step function be

$$f = 2\chi_{[1,3]} + 4\chi_{[0,2]} + \chi_{(1,4)} + 3\chi_{(2,6)} \tag{10}$$

This step function can be rewritten to only contain disjoint intervals using the construction above. So,

$$f = 4\chi_{[0,0]} + (2+4)\chi_{[1,1]} + (2+4+1)\chi_{[2,2]} + (2+1+3)\chi_{[3,3]} + 3\chi_{[4,4]} + 0\chi_{[6,6]} + 4\chi_{(0,1)} \tag{11}$$

$$+(2+4+1)\chi_{(1,2)} + (2+1+3)\chi_{(2,3)} + (1+3)\chi_{(3,4)} + 3\chi_{(4,6)}$$

Lastly, this step function can be simplified even further by combining adjacent intervals with the same coefficients to get the unique representation

$$f = 6\chi_{[1,1]} + 4\chi_{[0,1)} + 7\chi_{(1,2]} + 6\chi_{(2,3]} + 4\chi_{(3,4)} + 3\chi_{[4,6)} \tag{12}$$

Theorem 2.2. *The Lebesgue integral of a step function f is well-defined. In other words, if f has two different representations*

$$f = \sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x) = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x) \tag{13}$$

then we have that

$$\int_{\mathbb{R}} f dx = \sum_{i=1}^{n} k_i \cdot \mu(J_i) = \sum_{j=1}^{m} c_j \cdot \mu(I_j) \tag{14}$$

Proof. First, let's attempt to show that the integral of any step function f is equal to the integral of the equivalent representation of f that contains the least number of disjoint intervals shown above. To do this, let

$$f = \sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x) = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$$

where the second step function to the very right of the equality is the unique representation of f shown in Lemma 2.1 that is only defined on disjoint intervals and has the minimum number of intervals I_j in its representation. So, assume just as we did in Lemma 2.1, that the coefficients of the second step function are defined to be $c_s = \sum k_i$ for all i where $I_s \subseteq J_i$.

We also know from Definition 2.13 that $\int_{\mathbb{R}} f(x)dx = \sum_{i=1}^{n} k_i \cdot \mu(J_i)$. Since the set of intervals $\{I_j\}$ partition the set of intervals $\{J_i\}$, we know that summing over all intervals $I_j \subseteq J_i$

$$k_i \cdot \mu(J_i) = \sum k_i \cdot \mu(I_j)$$

Now, summing over all the J_i intervals corresponds to summing over the $r-1$ nontrivial intervals $I_j \subseteq J_i$ so

$$\sum_{i=1}^{n} k_i \cdot \mu(J_i) = \sum_{i=1}^{n} (\sum k_i \cdot \mu(I_j)) = \sum_{i=j}^{r-1} c_j \cdot \mu(I_j)$$

Lastly, we can also include all of the trivial intervals in our rightmost sum because they are defined to have Lebesgue measure zero since the Lebesgue measure of a finite collection of real numbers is 0. Finally,

$$\sum_{i=1}^{n} k_i \cdot \mu(J_i) = \sum_{i=j}^{m} c_j \cdot \mu(I_j)$$

Lastly, by Definition 2.13, we have that the integrals of the two step functions $\sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x)$ and $\sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$ are in fact the same and equal the Lebesgue integral of f.

Finally, for any step function representation $\sum_{i=1}^{n} k_i \cdot \chi_{J_i}(x)$ of f, we have that if $\sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$ is the simplified step function representation defined in Lemma 2.1, then

$$\int_{\mathbb{R}} f(x)dx = \sum_{i=1}^{n} k_i \cdot \mu(J_i) = \sum_{i=j}^{m} c_j \cdot \mu(I_j)$$

Now, this result shows us that if our function f has two different step function representations, then we can reduce both step functions to the unique step function on disjoint intervals defined in Lemma 2.1 such that the Lebesgue integral of the first two step functions is equal to the Lebesgue integral of the simplified step function. Therefore, the Lebesgue integral is well-defined on the set of step functions. □

Now we will show the space of step functions is closed under linear combinations (linear space).

Theorem 2.3. *Let f and g be any step functions and a, b be real constants. Then, the function $a \cdot f + b \cdot g$ is a step function and*

$$\int_{\mathbb{R}} (a \cdot f + b \cdot g)dx = a \int_{\mathbb{R}} f dx + b \int_{\mathbb{R}} g dx \tag{15}$$

Proof. Define $f = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$ and $g = \sum_{j=m+1}^{n} c_j \cdot \chi_{I_j}(x)$. Then,

$$a \cdot f + b \cdot g = \sum_{j=1}^{n} k_j \cdot \chi_{I_j}(x) \tag{16}$$

where $k_j = a \cdot c_j$ for $1 \leq j \leq m$ and $k_j = b \cdot c_j$ for $m+1 \leq j \leq n$. Therefore, $a \cdot f + b \cdot g$ is a step function by definition. Using Definition 2.13, we can calculate its integral to be

$$\int_{\mathbb{R}} (a \cdot f + b \cdot g)dx = \int_{\mathbb{R}} \sum_{j=1}^{n} k_j \cdot \chi_{I_j}(x)dx = \sum_{j=1}^{n} k_j \cdot \mu(I_j) = \sum_{j=1}^{n} a \cdot c_j \cdot \mu(I_j) + \sum_{j=1}^{n} b \cdot c_j \cdot \mu(I_j) = a \int_{\mathbb{R}} f dx + b \int_{\mathbb{R}} g dx$$

□

Theorem 2.4. *Let f and g be step functions with $f \geq g$ on all of $\mathbb{R}$. Then, $\int_{\mathbb{R}} f dx \geq \int_{\mathbb{R}} g dx$.*

Proof. We know by Theorem 2.3 that $f(x) - g(x)$ is a step function. So we can represent our step function as $f(x) - g(x) = \sum_{j=1}^{m} c_j \cdot \chi_{I_j}(x)$ where each of the c_j's is nonnegative since our step function is nonnegative. Therefore, we can write out by Definition 2.13 that

$$\int_{\mathbb{R}} (f - g)dx = \sum_{j=1}^{m} c_j \cdot \mu(I_j)$$

The sum to the right is also nonnegative because all the c_j's must be nonnegative and any measure μ must be nonnegative. Finally,

$$\int_{\mathbb{R}} f dx - \int_{\mathbb{R}} g dx = \int_{\mathbb{R}} (f - g)dx \geq 0$$

□

Now we will describe the L^0 space of functions which we briefly mentioned at the beginning of this section.

Definition 2.14 (L^0 Space)**.** Let $\{\phi_n(x)\}$ be a sequence of nondecreasing step functions that converges pointwise almost everywhere to a function $f(x)$ that is defined on all of $\mathbb{R}$. By nondecreasing, we imply that on all of the reals, $\phi_{n+1}(x) \geq \phi_n(x)$ and by pointwise convergence almost everywhere,

we mean that the sequence $\{\phi_n(x)\}$ converges to $f(x)$ at every point on the reals except on a set of Lebesgue measure zero. If these conditions are satisfied, then we define the integral of f as

$$\int_{\mathbb{R}} f dx = \lim_{n\to\infty} \int_{\mathbb{R}} \phi_n dx \tag{17}$$

For functions f that have a finite integral as defined above, we say f belongs to the space L^0.

Note that in the definition of the integral above, the limit of the integral of the step functions will always exist. This is because the sequence of step functions $\{\phi_n\}$ is nondecreasing so our limit of nondecreasing terms is either finite or unbounded. Now we will show that our definition for the integral of any function in L^0 is consistent.

Theorem 2.5. *Let $\{\phi_n\}$ and $\{\varphi_n\}$ be two nondecreasing sequences of step functions whose integrals are bounded and converge almost everywhere to a function f. Then,*

$$\lim_{n\to\infty} \int_{\mathbb{R}} \phi_n dx = \lim_{n\to\infty} \int_{\mathbb{R}} \varphi_n dx \tag{18}$$

Proof. Suppose f and g are in L^0 where $f \geq g$ almost everywhere. Assume $\{\phi_n\}$ is a sequence of nondecreasing step functions that tends to f and that $\{\varphi_n\}$ is a sequence of nondecreasing step functions that tends to g.

Then we know by Definition 2.14 that $\lim_{n\to\infty} \int_{\mathbb{R}} \phi_n dx = f$ and $\lim_{n\to\infty} \int_{\mathbb{R}} \varphi_n dx = g$. Since limits preserve inequalities, we know that

$$\int_{\mathbb{R}} f dx = \lim_{n\to\infty} \int_{\mathbb{R}} \phi_n dx \geq \lim_{n\to\infty} \int_{\mathbb{R}} \varphi_n dx = \int_{\mathbb{R}} g dx$$

Now, if $f = g$ almost everywhere, then we get from the above inequality that

$$\lim_{n\to\infty} \int_{\mathbb{R}} \phi_n dx = \lim_{n\to\infty} \int_{\mathbb{R}} \varphi_n dx$$

and we are done. □

Now that we have defined the space L^0 and shown that it is consistent, we will provide an argument as to why the space L^0 does not satisfy the properties we would want a space of integrable functions to have. For example, it can be shown that the a sequence of pointwise convergent functions in L^0 does not necessarily converge to another function in L^0. Even worse, the space of L^0 functions is not linear. In the next argument, we will generate two functions in L^0 such that their linear composition is not in L^0 [2].

Let f be the characteristic function of the interval $(0,1)$ so $f = \chi_{(0,1)}$. Evidently, f is a step function so $f \in L^0$. Define the function g to be the limit of the characteristic functions $\{\bigcup_{j=1}^n I_j\}$ where for all the ordered rational numbers a_j in $(0,1)$, we have that $I_j = (a_j - \frac{1}{2^{j+2}}, a_j + \frac{1}{2^{j+2}})$. Therefore, $g = \lim_{n\to\infty} \chi_{(\bigcup_{j=1}^n I_j)}$. Note that g is defined as a limit of nondecreasing step functions.

We know that since f is a simple step function, $\int_{\mathbb{R}} f dx = 1$. Also, we have seen in one of the earlier examples that $\int_{\mathbb{R}} \chi_{(\bigcup_{j=1}^\infty I_j)} dx = \sum_{j=1}^\infty \mu(I_j) = \frac{1}{2}$. Therefore,

$$\int_{\mathbb{R}} g dx = \lim_{n\to\infty} \int_{\mathbb{R}} \chi_{(\bigcup_{j=1}^n I_j)} dx \leq \lim_{n\to\infty} \sum_{j=1}^n \chi_{I_j} dx = \sum_{j=1}^\infty \mu(I_j) = \frac{1}{2} \tag{19}$$

Finally, g is a limit of nondecreasing step functions and its integral as defined by Definition 2.14 is bounded so $g \in L^0$. If L^0 where a linear space, then this would mean that

$$\int_{\mathbb{R}} (f-g) dx = \int_{\mathbb{R}} f dx - \int_{\mathbb{R}} g dx \geq 1 - \frac{1}{2} = \frac{1}{2} \tag{20}$$

In the above equation, we are allowed to split up the integrand as we did with the first equality by the results proven in Theorem 2.3. These results still hold under the limit because the limit is a linear operator.

If $f-g$ is in L^0, then Definition 2.14 requires a sequence of nondecreasing step functions $\{\phi_n\}$ that converges pointwise almost everywhere to $f-g$. But this means that every ϕ_n is less than or equal to $f-g$. So any step function ϕ_n in our sequence would be 0 everywhere except on the points outside of $\bigcup_{j=1}^{n} I_j$ where it may equal 0 or 1.

Now, let $x \in (0,1)$ be a value where $f(x)-g(x)=1$ and take a positive length interval I that contains x. This interval must contain a rational number a_j and therefore, I contains a nontrivial interval $A = I_j \cap I \cap (0,1)$. Furthermore, we know in this interval A that $f-g=1-1=0$ is identically zero. We know that by Lemma 2.1, the unique step function representation of ϕ_n will contain the term $c \cdot \chi_I$ for some constant c. Since $f-g$ is identically zero on the interval A, we must have that $c \leq 0$. Therefore, $\phi_n \leq 0$ for all $n = 1, 2, ...$ on the interval $(0,1)$.

But this implies that $\int_{\mathbb{R}} f dx - \int_{\mathbb{R}} g dx = \lim_{n\to\infty} \int_{\mathbb{R}} \phi_n dx \leq \lim_{n\to\infty} 0 = 0$ which is a contradiction since we assumed that $f-g$ is in L^0 and that $\int_{\mathbb{R}} (f-g)dx \geq \frac{1}{2}$. Therefore, $f-g$ cannot be in L^0.

So, we have just shown that the L^0 space doesn't satisfy one of the most fundamental properties we would want our space of integrable functions to have. Nevertheless, we can use the L^0 space to define a space of functions called L^1 which is equivalent to the space of Lebesgue integrable functions.

Definition 2.15 (L^1 Space)**.** The space L^1 of Lebesgue integrable functions consists of any function f of the form $f = g - h$ where g, h are in L^0. Furthermore, the Lebesgue integral of f is defined as

$$\int_{\mathbb{R}} f dx = \int_{\mathbb{R}} g dx - \int_{\mathbb{R}} h dx \tag{21}$$

Note that by the above definition, setting $h = 0$ we have that any function g in L^0 is also in L^1. This definition of L^1 fixes the problem of nonlinearity that we discovered in the L^0 space.

Theorem 2.6. *Let f and u be two functions in L^1 and let a, b be arbitrary constants. Then, $a \cdot f + b \cdot u$ is in L^1 and*

$$\int_{\mathbb{R}} (af + bu)dx = a \int_{\mathbb{R}} f dx + b \int_{\mathbb{R}} u dx \tag{22}$$

Proof. Since f and g are in L^1, we have by Definition 2.16 that $f = g - h$ and $u = v - w$ where g, h, v, w are all in L^0. Lets first assume that a and b are both nonnegative. Then,

$$af + bu = ag - ah + bv - bw = (ag + bv) - (ah + bw)$$

We want to show that $ag + bv$ and $ah + bw$ are both in L^0. To do this, let's first prove that for functions s and t in L^0 and nonnegative constants a and b, we have that $as + bt$ is in L^0. Let the sequence of step functions $\{\phi_n\}$ converge almost everywhere to s and let the sequence of step functions $\{\varphi_n\}$ converge almost everywhere to t. Then, by the linearity of limits, $\{a\phi_n + b\varphi_n\}$ is a sequence of step functions that converges almost everywhere to $as + bt$. Therefore, by Definition 2.14, $as + bt$ is in L^0. Also, we showed in Theorem 2.3 that integration of a step function is linear and limit operations are also linear so

$$\int_{\mathbb{R}} (as + bt)dx = \lim_{n\to\infty} \int_{\mathbb{R}} (a\phi_n + b\varphi_n)dx = \lim_{n\to\infty} a \int_{\mathbb{R}} \phi_n dx + \lim_{n\to\infty} b \int_{\mathbb{R}} \varphi_n dx$$

$$= a \int_{\mathbb{R}} s dx + b \int_{\mathbb{R}} t dx$$

So finally, in our original statement, since a and b are both nonnegative, then $ag+bv$ and $ah+bw$ are both in L^0. This implies by Definition 2.15 that $af + bu$ is in L^1.

Now, there are three more cases to examine.

If a is nonnegative and b is nonpositive, then we can write

$$af + bu = ag - ah + bv - bw = (ag + (-b)w) - (ah + (-b)v)$$

If a is nonpositive and b is nonnegative, then we can write

$$af + bu = ag - ah + bv - bw = ((-a)h + bv) - ((-a)g + bw)$$

If a is nonpositive and b is nonpositive, then we can write

$$af + bu = ag - ah + bv - bw = ((-a)h + (-b)w) - ((-a)g + (-b)v)$$

Therefore, in each one of the cases above, we can reduce our problem to the first case where a and b are nonnegative and can proceed with the same proof. □

Now we will prove by a quick application of the previous theorem that our definition of Lebesgue integration in L^1 is consistent.

Corollary 2.6.1. Assume that f has two different representations in L^1 and $f = g - h = v - w$ where g, h, v, w are all in L^0. Then, $\int_{\mathbb{R}} f dx = \int_{\mathbb{R}} g dx - \int_{\mathbb{R}} h dx = \int_{\mathbb{R}} v dx - \int_{\mathbb{R}} w dx$.

Proof. We know that $f = g - h = v - w$ so $g + w = v + h$. We know by Definition 2.15 that $g + w$ and $v + h$ are in L^1 and by the linearity of the Lebesgue integral, $\int_{\mathbb{R}} g dx + \int_{\mathbb{R}} w dx = \int_{\mathbb{R}} v dx + \int_{\mathbb{R}} h dx$. □

Theorem 2.7. *If f is in L^1 and $f \geq 0$, then $\int_{\mathbb{R}} f dx \geq 0$.*

Proof. We can write $f = g - h$ where g and h are in L^0. Since $f \geq 0$, we know that $g \geq h$ almost everywhere. Now, by Definition 2.15, we have that $\int_{\mathbb{R}} f dx = \int_{\mathbb{R}} g dx - \int_{\mathbb{R}} h dx$ and since $g \geq h$ we have by Theorem 2.4 and by the fact that limits preserve inequalites (with the exception of strictness) that $\int_{\mathbb{R}} g dx \geq \int_{\mathbb{R}} h dx$. Finally, we have that $\int_{\mathbb{R}} f dx \geq 0$. □

3 Some Basics of Probability and Chebyshev's Inequality

3.1 Probability Space

Now that we have had an introduction to measure theory, we are ready to talk about the idea of a probability space. Many of these definitions are taken from the Cambridge University class notes of J. R. Norris [4].

Definition 3.1 (Probability Space)**.** A probability space is a triplet $(\Omega, \mathcal{F}, P)$ that describes a model for a class of real-world situations:

- Ω represents the sample space which is a non-empty set of all possible outcomes of the model being executed.
- $\mathcal{F}$ is a σ-algebra on our sample space where $\Omega \in \mathcal{F}$. Note that $\mathcal{F}$ inherits all the properties of a σ-algebra described in the previous section.
- P is called a probability measure if
  - It is a measure that maps any event in $\mathcal{F}$ to its probability from 0 to 1 and the total measure of the sample space is 1. In other words, $P : \mathcal{F} \to [0, 1]$ and $P(\Omega) = 1$.
  - It is countably additive. So if $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Let's attempt to describe the triplet $(\Omega, \mathcal{F}, P)$ in a way that agrees with our intuitive understanding of probability. A probability space is a structure that presents a mathematical model describing a real-world system. This real-world system must have a set of outcomes Ω and the probability space attempts to assign a probability by the measure P to all measurable subsets of outcomes (events) which are elements of $\mathcal{F}$.

As an example, take a discrete model to be the process of flipping a coin twice and recording which way it lands (H or T) on both turns [1]. In this case, our sample space is the set $\Omega =$

$\{HH, HT, TH, TT\}$. One possible element of our σ-algebra could be the event $E = \{HH, TT\}$ in which we get all heads or all tails on both flips. Then, we could define a probability measure

$$P(A) = \sum_{x \in A} \frac{1}{4} = \frac{|A|}{4} \tag{23}$$

where $|A|$ is the number of elements in the set of outcomes A. Then, in this case, we get that $P(E) = \frac{1}{2}$ which is exactly what we expected it would be. The reason our probability for E is appropriate is because we chose a reasonable probability measure P that we assumed worked well for the model. In general, we could have chosen any arbitrary P as long as it satisfied the properties of a probability measure.

Of course, the same definition of probability spaces also applies to processes with a continuous sample space. For example, as described in [1], we could take the example of spinning a pointer on a dial where the pointer can take on any angle value on the half-open set $\Omega = [0, 2\pi)$.

As a reasonable assumption, we may want our probability measure to have the following properties

$$P(I) = \frac{\text{length of } I}{2\pi}, \quad P(E) = 1 \tag{24}$$

Now the question comes to how we will choose a set of events $\mathcal{F}$ for the domain of our probability measure P. Is it possible to extend the domain of P to the class of all subsets of Ω? In general, the answer is no as this set of all subsets does not satisfy the properties of a σ-algebra. This can be rigorously proven if we are allowed to use the axiom of choice and the continuum hypothesis.

What probability theory allows us to say in general, is that our probability measure P can be extended to all *Lebesgue measurable* sets of Ω. This is precisely why a probability space specifies that P must be a measure which means that its domain contains only measurable sets. In other words, we cannot determine the probability of an event in $\mathcal{F}$ if that set is non-measurable.

3.2 Random Variables, Probability Mass Functions, and Probability Density Functions

Random variables and probability densities are two concepts in probability that are closely related to each other and are often presented together. They come in two flavors, discrete and continuous. In its simplest form, a random variable is a variable whose value depends on the outcomes of a system. For example, if we go back to the coin toss example, we could define a discrete real random variable X to be 5 if the coin fell on heads and -10 if the coin fell on tails. We could also make an educated guess that if the tossed coin is fair, then the probability of attaining a heads or a tails is $\frac{1}{2}$. Assigning a probability to an outcome of a discrete random variable is the purpose of the *probability mass function*. More formally, we could write that for a sample space $\Omega = \{\text{heads}, \text{tails}\}$

$$X(\omega) = \begin{cases} 5, & \omega = \text{heads} \\ -10, & \omega = \text{tails} \end{cases} \tag{25}$$

and we could define the probability mass function for X as

$$p_X(x) = \begin{cases} \frac{1}{2}, & x = 5 \\ \\ \frac{1}{2}, & x = -10 \end{cases} \tag{26}$$

How would we define the terms above if we were to apply them to a continuous system such as the spinner we discussed in the previous section? We have already noted that the sample space of the angle of the spinner is $\Omega = [0, 2\pi)$. Therefore, we could create a random variable X that attains a value equal to the angle of the spinner when it is spun. Assigning a probability that the spinner angle lands in a certain subset of $[0, 2\pi)$ is a bit harder. We can no longer ask what the probability is of the spinner landing exactly at the angle π. The probability of our spinner angle equaling exactly some value to infinite precision is zero. We must instead say that our random variable X has a *probability*

density function which defines the infinitesimal probability that X attains any given value. The use of the probability density function is in computing the probability that our random variable attains a value within some range (or within a set that has a nonzero measure). For example, we may want for our probability density function to have the following property as shown in equation (24). That is, we could give the probability that our spinner angle random variable X is within some interval $I = [\alpha, \beta)$ but the probability that our spinner angle is exactly any point within that interval is 0. We could define X and its probability density function as follows

$$X(\alpha) = \alpha \tag{27}$$

$$f_X(x) = \frac{1}{2\pi} \tag{28}$$

In order to obtain the probability that X is within some subset of our sample space, we will Lebesgue integrate the probability density function for that random variable over the subset and its value (normalized real number from 0 to 1). This will give us the probability that the random variable X takes on any value in that set. We will go into more detail on this in the next section.

Definition 3.2 (Random Variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(E, \mathcal{E})$ be a measurable space. We call X an $(E, \mathcal{E})$-valued random variable if $X : \Omega \to E$ is $(\mathcal{F}, \mathcal{E})$-measurable. $(\mathcal{F}, \mathcal{E})$-measurable means that for every subset $S \in \mathcal{E}$, its preimage $X^{-1}(S) \in \mathcal{F}$.

Simply stated, X takes on values in E for all the outcomes in its sample space Ω. This definition for a random variable establishes that X is a measurable map that maps measurable subsets of outcomes $\mathcal{F}$ (events) in its probability space to measurable subsets in its range $\mathcal{E}$ (values that the variable takes on). $(E, \mathcal{E})$ must be a measurable space because we only wish to find probabilities that our variable X lies in measurable subsets of E. The additional assumption that X must be $(\mathcal{F}, \mathcal{E})$-measurable enforces that any measurable subset of E must have a measure determined by the preimage of X. In other words, the probability that X lies in a measurable subset of E is well-defined and will always exist.

A real-valued random variable is a random variable X where $X : \Omega \to \mathbb{R}$.

Now, for the purposes of this paper, we will define the concept of the probability density function without any additional measure-theoretic concepts and with a few more restrictive properties so that we can more readily deal with real-valued random variables and their densities.

Definition 3.3 (Probability Density Function)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(E, \mathcal{E})$ be a measurable space. Let X be a $(E, \mathcal{E})$-valued random variable where $X : \Omega \to E$. Then, we say that X admits a real-valued probability density function $f_X : E \to [0, \infty)$ if it is a Lebesgue integrable function and it satisfies the following three properties:

- $f_X(x) \geq 0$ for all $x \in E$
- For any measurable set $A \in \mathcal{E}$, we assume that we can write P as an integral in terms of f_X in the following way
$$P(\{\omega \in \Omega : X(\omega) \in A\}) = \int_A f_X d\mu \tag{29}$$
- The integral of the probability density function over the entire range of X is 1
$$P(\Omega) = \int_E f_X d\mu = 1 \tag{30}$$

Note that by the above definition, the probability measure P still satisfies the properties of a probability measure in a probability space $(\Omega, \mathcal{F}, P)$. The first property ensures that our probability measure will never be negative assuming that f is continuous. The second property makes sure that P is countably additive since the Lebesgue integral is countably additive. The third property makes

sure that the probability measure P never exceeds 1 and that the probability measure of the entire sample space is 1.

The probability density function f_X is sometimes called the *Radon-Nikodym derivative* of the probability measure P and it is denoted by $f_X = \frac{dX_*P}{d\mu}$.

3.3 Expected Value, Variance, and Important Probability Distributions

Though in the previous sections we mentioned discrete and continuous examples of various probabilistic phenomena, we will now treat both discrete and continuous variables in the same way due to a unifying interpretation that we will state below. The reason for this is that the discrete cases of these definitions often involve summation rather than Lebesgue integration and use a measure known as the counting measure (simply the number of elements in a set) rather than the Lebesgue measure. In order to prove all of our results in more generality, we will try to not distinguish between discrete and continuous cases of the same result. Furthermore, we will restrict all of our random variables to be real-valued. We will commonly refer to a random variable X together with its probability density function f_X.

Before we can talk about what a probability distribution is, we must define the cumulative distribution function which is a function on the reals.

Definition 3.4 (Cumulative Distribution Function)**.** The cumulative distribution function for a real valued random variable $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, P)$ is the function given by

$$F_X(x) = P(X \leq x) \tag{31}$$

For all real continuous random variables, it can be defined as the integral of the probability density function over an interval.

Definition 3.5. For a real continuous random variables X with probability density f_X, the cumulative distribution function is

$$F_X(a) = \int_{-\infty}^{a} f_X(x)dx \tag{32}$$

For all real discrete random variables, the cumulative distribution function can be defined as a sum of step functions weighted by the respective probability at that point.

Definition 3.6. For a real discrete random variables X with probability mass function p_X and range $R_X = \{x_1, x_2, x_3, ...\}$ (where the x_k are real), the cumulative distribution function is

$$F_X(x) = \sum_{x_k \in R_X} p_X(x_k)u(x - x_k) \tag{33}$$

where the step function u is defined to be

$$u(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{34}$$

Now that we have described the two different types of random variables and their cumulative distribution functions, we must state a very interesting unification of these two definitions. If one is to look at probability mass functions more closely, it becomes very clear that any probability mass function can be written as a sum of Dirac delta functions weighted by the probabilities assigned to those values by the probability mass function of that random variable. This approach can be useful if a probability density function happens to be both continuous in one region and discrete in another. Therefore, any discrete probability mass function can be treated as a generalized probability density function. We will not attempt to prove this result in this paper.

The expected value for a real-valued, random variable X is a weighted sum or integral of the values of X by the probability that it attains those values.

Definition 3.7 (Expected Value)**.** Let $X : \Omega \to \mathbb{R}$ be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$ with probability density function f_X. The expected value of a random variable X is defined to be

$$E[X] = \int_{\omega \in \Omega} X(\omega) dP(\omega) = \int_{\mathbb{R}} x f_X(x) dx \tag{35}$$

As the reader may be familiar with, we will commonly refer to the expected value as the average value or mean. One important property of the expected value is that it is linear. This is commonly referred to as the *linearity of expectation* and we will use it in one of the later examples when we want to compute the variance of a random variable.

The variance of a real-valued, random variable X is defined as the expected value of the squared deviation of X from its mean. In general, it is a value that tells us how "spread out" the values of X are from their mean.

Definition 3.8 (Variance)**.** Let $X : \Omega \to \mathbb{R}$ be a continuous real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$ with probability density function f_X. Let the mean of X be $m = E[X]$. The variance of X is defined to be

$$Var(X) = E[(X - m)^2] = \int_{\omega \in \Omega} (X(\omega) - m)^2 dP(\omega) = \int_{\mathbb{R}} (x - m)^2 f_X(x) dx \tag{36}$$

We will commonly denote the variance as σ^2 which means the *standard deviation* squared.

Definition 3.9 (Standard Deviation)**.** Let $X : \Omega \to \mathbb{R}$ be a continuous real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$ with probability density function f_X. Let the variance of X be $Var(X)$. Then, the standard deviation σ is defined to be $\sigma = \sqrt{Var(X)}$.

Note that in the above definition, we are taking the square root of variance. This is justified by the fact that the variance of a random variable can never be negative as it is the expected value of squares.

Now, that we have defined probability density functions, cumulative distribution functions, and mentioned how we could unify both the discrete and continuous cases using the Dirac delta function, we can easily talk about a *probability distribution*. Although the terms probability distribution and cumulative distribution function are used interchangeably in many different sources, we will make sure to define a probability distribution as a pair of both a specific probability density function along with its cumulative distribution.

For example, the normal distribution has a probability density of the form

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{37}$$

where μ is the mean of the distribution, σ is its standard deviation, and σ^2 is its variance.

By integrating the density function of the normal distribution, we get its cumulative distribution.

$$\frac{1}{2}[1 + erf(\frac{x - \mu}{\sigma\sqrt{2}})] \tag{38}$$

The derivation for the cumulative distribution function above is a standard process of integration which we will not go into here and it involves defining the error function (erf). Nevertheless, we will now introduce the binomial distribution which is a discrete distribution. Unlike the normal distribution, the binomial distribution does not have parameters for mean and variance and we must calculate them directly. The binomial distribution represents the probability of getting j successes out of n total experiments where the probability of success in each experiment is x.

For notational purposes, we will write the probability mass function of the binomial distribution as

$$b(n, x; j) = \binom{n}{j} x^j (1 - x)^{n-j} \tag{39}$$

The cumulative distribution of the binomial distribution can be written as a simple sum

$$F(n,x;j) = \sum_{i=1}^{\lfloor j \rfloor} \binom{n}{i} x^i (1-x)^{n-i} \tag{40}$$

The mean of a binomially distributed random variable X where the binomial distribution has parameters n, x is

$$E[X] = \sum_{j=1}^{n} j \cdot b(n,x;j) = \sum_{j=1}^{n} j \cdot \binom{n}{j} x^j (1-x)^{n-j}$$

$$= nx \sum_{j=1}^{n} \frac{(n-1)!}{(n-j)!j!} x^{j-1} (1-x)^{(n-1)-(j-1)} = nx \sum_{j=1}^{n} \frac{(n-1)!}{((n-1)-(j-1))!j!} x^{j-1} (1-x)^{(n-1)-(j-1)}$$

$$= nx \sum_{j=1}^{n} \binom{n-1}{j-1} x^{j-1} (1-x)^{(n-1)-(j-1)}$$

Setting $l = j-1$ and $m = n-1$ we have that

$$E[X] = nx \sum_{l=0}^{m} \binom{m}{l} x^l (1-x)^{m-l} = nx(x + (1-x))^m = nx \tag{41}$$

Now let's compute the variance of the binomial distribution. By linearity of expectation,

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[2XE[X]] + E[E[X]^2] \tag{42}$$

$$= E[X^2] - 2E[X]^2 + E[X]^2 = E[X^2] - E[X]^2$$

The last equality in the above formula can be derived from the fact that the expected value of a constant is equal to that constant.

Using the above formula, we can write

$$Var(X) = E[X^2] - E[X]^2 = E[X^2] - E[X] + E[X] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2 \tag{43}$$

So, for the binomial distribution,

$$E[X(X-1)] = \sum_{j=2}^{n} j(j-1) \cdot b(n,x;j) = \sum_{j=2}^{n} j(j-1) \cdot \frac{n!}{j!(n-j)!} x^j (1-x)^{n-j}$$

$$= \sum_{j=2}^{n} \frac{n!}{(j-2)!(n-j)!} x^j (1-x)^{n-j}$$

Letting $k = j-2$ we have

$$E[X(X-1)] = \sum_{k=0}^{n-2} \frac{n!}{k!(n-2-k)!} x^{k+2} (1-x)^{n-2-k} = n(n-1)x^2 \sum_{k=0}^{n-2} \frac{(n-2)!}{k!(n-2-k)!} x^{k+2} (1-x)^{n-2-k}$$

$$= n(n-1)x^2 \sum_{k=0}^{n-2} b(n-2,x;k) = n(n-1)x^2$$

Finally, the variance of the binomial distribution is

$$Var(X) = E[X(X-1)] + E[X] - E[X]^2 = n(n-1)x^2 + nx - (nx)^2 \tag{44}$$

$$= n^2x^2 - nx^2 + nx - n^2x^2 = nx(1-x)$$

3.4 Chebyshev's Inequality

Now we will state Chebyshev's Inequality and will prove it in its most general form using Lebesgue integration [4].

Theorem 3.1 (General Chebyshev's Inequality)**.** *Let (W, σ, μ) be a measure space and let f be a real-valued measurable function defined on W. μ is the Lebesgue measure. Also, let g be a real-valued measurable function that is nonnegative and nondecreasing on the range of f. Then, for any real number $t > 0$ and $0 < p < \infty$, we have that*

$$\mu(\{x \in W : f(x) \geq t\}) \leq \frac{1}{g(t)} \int_W g(f(x)) d\mu(x) \tag{45}$$

Proof. Let t be fixed and define $A_t = \{x \in W : f(x) \geq t\}$. Define the characteristic function χ_{A_t} on the set A_t. Since g is nondecreasing and it is nonnegative on the range of f, we must have that

$$0 \leq g(t)\chi_{A_t} \leq g(f(x))\chi_{A_t}$$

Note that the second inequality is true because $f(x) \geq t$ on A_t and g is nondecreasing. On points x outside of A_t, all three parts of the inequality above are equivalently zero. Now, using Lebesgue integration to integrate over W, we get that

$$g(t)\mu(A_t) = g(t) \int_W \chi_{A_t} d\mu = \int_W g(t)\chi_{A_t} d\mu$$

By Theorem 2.6 and 2.7, we know that Lebesgue integration preserves inequalities so

$$\int_W g(t)\chi_{A_t} d\mu \leq \int_W g(f(x))\chi_{A_t} d\mu = \int_{A_t} g(f(x)) d\mu \leq \int_W g(f(x)) d\mu$$

The last inequality in the formula above holds because g is nonnegative everywhere.
Finally, we have that

$$\mu(A_t) \leq \frac{1}{g(t)} \int_W g(f(x)) d\mu$$

□

In reality, Chebyshev's Inequality is rarely presented in the form above and is typically shown as defined below.

Theorem 3.2 (Chebyshev's Inequality)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. Suppose X has finite expected value m and finite nonzero variance σ^2. Then, for any real number $k > 0$,*

$$P(|X - m| \geq k\sigma) \leq \frac{1}{k^2} \tag{46}$$

Proof. Our general measure-theoretic description of Chebyshev's Inequality can be reduced to the one defined here. We know from Theorem 3.1 that for a real-valued measurable function f defined on $\mathbb{R}$ and real-valued measurable function g that is nonnegative and nondecreasing on the range of f we have that

$$P(\{\omega \in \Omega : f(X(\omega)) \geq t\}) \leq \frac{1}{g(t)} \int_{\omega \in \Omega} g(f(\omega)) dP(w)$$

for any real number $t > 0$ and $0 < p < \infty$.

Now, we can let $g(t) = t^2$ if $x \geq t$ and 0 otherwise. Also, we can substitute $|X(\omega) - m|$ for f and let $t = k\sigma$ in the above inequality. Then we will get that

$$P(\{\omega \in \Omega : |X(w) - m| \geq k\sigma\}) \leq \frac{1}{k^2\sigma^2} \int_{\omega \in \Omega} |X(\omega) - m|^2 dP(w)$$

$$P(|X-m| \geq k\sigma) \leq \frac{1}{k^2\sigma^2}\int_{\omega\in\Omega}(X(\omega)-m)^2 dP(\omega)$$

But the integral in the rightmost expression is just equal to $Var(X) = \sigma^2$ so finally, we get that

$$P(|X-m| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

□

4 Approximation of Functions by Polynomials

Now that we have built up all the necessary probabilistic theory, we can give a concise and elegant proof of the Weierstrass Approximation Theorem [3].

4.1 Bernstein's Proof of the Weierstrass Approximation Theorem

Theorem 4.1 (Weierstrass Approximation Theorem)**.** *If f is a continuous real-valued function on the interval $[0,1]$, then for arbitrary $\epsilon > 0$, there exists a polynomial function p such that $\max|f(x) - p(x)| < \epsilon$ where $0 \leq x \leq 1$.*

Definition 4.1. The Bernstein polynomials are defined as

$$B_n(f;x) = \sum_{j=0}^{n} f(\frac{j}{n})\binom{n}{j}x^j(1-x)^{n-j}, \quad n = 0,1,2,... \tag{47}$$

Perhaps the main reason why Bernstein's proof is so widely supported is due to the fact that it is constructive and allows us to actually find a sequence of polynomials which uniformly converge to any function on the interval. The proof is given below and it is a reproduction of Bernstein's Proof as shown in [3].

Proof. Let f be a continuous real valued function on the closed interval $[0,1]$. Also, let $F_{n,x}$ be a sequence of families of random variables that admit a probability density function which is a binomial distribution where $0 \leq x \leq 1$ is the probability of success and $n = 0,1,2,...$ is the total number of trials. In particular, define $F_{n,x}$ to have a value of $f(\frac{j}{n})$ if j successes occur.

Then, the expected value of $F_{n,x}$ can be calculated through a finite sum

$$E[F_{n,x}] = \sum_{j=0}^{n} f(\frac{j}{n})b(n,x;j) = \sum_{j=0}^{n} f(\frac{j}{n})\binom{n}{j}x^j(1-x)^{n-j} = B_n(f;x) \tag{48}$$

Our next and final step is to prove that $E[F_{n,x}]$ converges uniformly to f in n on the closed unit interval. First note that since f is continuous on the compact set $[0,1]$, it is bounded and uniformly continuous. Boundedness of f gives us that $|f(x)| \leq M$ on $[0,1]$ where M is a real bound for our function. From this we have by the Triangle Inequality that for any two points $x, y \in [0,1]$, $|f(x) - f(y)| \leq 2M$.

Uniform continuity gives us that for all $\epsilon > 0$, there exists a $\delta > 0$ such that if $|x-y| < \delta$, then $|f(x) - f(y)| < \frac{\epsilon}{2}$.

For reasons that will soon become clear, let us choose an integer k such that $\frac{2M}{k^2} < \frac{\epsilon}{2}$ and choose positive integer N such that $\frac{k}{2\sqrt{N}} < \delta$. Then, for all $n \geq N$ we have that

$$|f(x) - B_n(f;x)| = |f(x) - E[F_{n,x}]| = \left|f(x) - \sum_{j=0}^{n} f(\frac{j}{n})b(n,x;j)\right| = \left|f(x)(x+(1-x))^n - \sum_{j=0}^{n} f(\frac{j}{n})b(n,x;j)\right|$$

$$= \left| f(x) \sum_{j=0}^{n} b(n,x;j) - \sum_{j=0}^{n} f(\frac{j}{n}) b(n,x;j) \right| \leq \sum_{j=0}^{n} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j)$$

$$= \sum_{\left|\frac{j}{n}-x\right| < \frac{k}{2\sqrt{n}}}^{n} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j) + \sum_{\left|\frac{j}{n}-x\right| \geq \frac{k}{2\sqrt{n}}}^{n} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j)$$

Since we chose $\frac{k}{2\sqrt{N}} < \delta$, we have that

$$\sum_{\left|\frac{j}{n}-x\right| < \frac{k}{2\sqrt{n}}}^{n} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j) \leq \sum_{\left|\frac{j}{n}-x\right| < \delta} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j)$$

By boundedness, we have that $\left|f(x) - f(\frac{j}{n})\right| \leq 2M$ so

$$\sum_{\left|\frac{j}{n}-x\right| \geq \frac{k}{2\sqrt{n}}}^{n} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j) \leq 2MPr(\left| \frac{j}{n} - x \right| \geq \frac{k}{2\sqrt{n}})$$

Therefore,

$$|f(x) - B_n(f;x)| \leq \sum_{\left|\frac{j}{n}-x\right| < \delta} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j) + 2MPr(\left| \frac{j}{n} - x \right| \geq \frac{k}{2\sqrt{n}})$$

Let's bound the first summand in the inequality above. Now, by uniform continuity $\left|f(x) - f(\frac{j}{n})\right| < \frac{\epsilon}{2}$ and since $\sum_{j=0}^{n} b(n,x;j) = (1 + (1-x))^n = 1$, we have that

$$\sum_{\left|\frac{j}{n}-x\right| < \delta} \left| f(x) - f(\frac{j}{n}) \right| b(n,x;j) \leq \frac{\epsilon}{2} \sum_{\left|\frac{j}{n}-x\right| < \delta} b(n,x;j) \leq \frac{\epsilon}{2}$$

Bounding the second summand involves the application of Chebyshev's Inequality. First, note that the maximum of $x(1-x)$ occurs at $x = \frac{1}{2}$ where $x(1-x) = \frac{1}{4}$. So we have that $\left|\frac{j}{n} - x\right| \geq \frac{k}{2\sqrt{n}} \geq k\sqrt{\frac{x(1-x)}{n}}$.

Therefore, we get that $|j - nx| \geq k\sqrt{nx(1-x)}$. So,

$$Pr(\left| \frac{j}{n} - x \right| \geq \frac{k}{2\sqrt{n}}) \leq Pr(|j - nx| \geq k\sqrt{nx(1-x)})$$

By the results proved in the previous section, we know that the mean of the binomial distribution $b(n,x;j)$ is $\mu = nx$ and its standard deviation is $\sigma = \sqrt{nx(1-x)}$ so applying Chebyshev's Inequality,

$$Pr(|j - nx| \geq k\sqrt{nx(1-x)}) \leq \frac{1}{k^2}$$

Finally, this result allows us to bound our second summand and our initial choice of k specified that $\frac{2M}{k^2} < \frac{\epsilon}{2}$ so

$$2MPr(\left| \frac{j}{n} - x \right| \geq \frac{k}{2\sqrt{n}}) \leq \frac{2M}{k^2} \leq \frac{\epsilon}{2}$$

Therefore, we have the result that

$$|f(x) - B_n(f;x)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \tag{49}$$

and we have successfully shown that $|f(x) - B_n(f;x)| < \epsilon$ for all $x \in [0,1]$ and $n \geq N$. Finally, this means that the Bernstein polynomials $B_n(f;x)$ converge uniformly to $f(x)$ on the closed unit interval $[0,1]$ which is exactly what we sought to prove.

□

5 Conclusion

In this paper, we introduced several important theories that we attempted to make understandable at the undergraduate level. We began with basic definitions of measure theory and then tied these concepts into probability theory to define some widely used probabilistic terms such as random variables and distributions. In order to discuss, probability density functions, mean, and variance, we needed to introduce the Lebesgue integral. In order to define it in a simple and understandable way for an undergraduate student, we used the Daniell-Riesz approach. The measure theoretic and probabilistic definitions we wrote out were in turn, used to prove the general form of Chebyshev's Inequality. Finally, we used this inequality in Bernstein's proof of the Weierstrass Approximation Theorem to prove that any continuous function on the unit interval could be approximated by a uniformly convergent sequence of polynomials.

One question that arises from the previous section is if the Bernstein polynomials can approximate any continuous function, then why are they not used in practice? The answer lies in not how close these polynomials approximate any continuous function, but how fast. The Bernstein polynomials have a significant drawback in the number of iterations it takes (how high of an n we have to take) to approximate certain functions well. It can be proven that to get an error term difference between a sequence of Bernstein polynomials and a continuous function f to be better than $1/n$, we must have that $f(x) = ax + b$ is a linear function. In general, Bernstein's polynomials are used for their theoretical implications and are rarely used for computational purposes.

References

[1] Botts, T. (2007). Probability Theory and the Lebesgue Integral. In Ross P. (Author) & Alexanderson G. (Ed.), *The Harmony of the World: 75 Years of Mathematics Magazine* (pp. 123-128). Mathematical Association of America. Retrieved from http://www.jstor.org/stable/10.4169/j.ctt13x0nm2.24

[2] Johnston, W. (2015). *Lebesgue Integral for Undergraduates.* (pp. 18-22, 26-34, 49-57). Washington DC: MAA Press.

[3] Kenneth M. Levasseur, *The American Mathematical Monthly* Vol. 91, No. 4 (Apr., 1984), (pp. 249-250).
Retrieved from http://ocw.nctu.edu.tw/course/fourier/supplement/Probabilistic%20Proof.pdf

[4] Norris, J. R. *Probability and Measure.* (n.d.), (pp. 1-30). Cambridge University. Retrieved from http://www.statslab.cam.ac.uk/ james/Lectures/pm.pdf