Expectation Values

Michael Betancourt

August 2023

Table of contents

In Chapter Three we defined measures, and probability distributions as a special case, as mappings from measurable subsets to allocated measures. These subset allocations, however, also induce a somewhat surprising but incredibly powerful mapping from real-valued functions to single real number *expectation value*. This *expectation* operation summarizes the interaction between a measure and a given function, allowing us to use one to learn about the other.

We will being our exploration of expectation values with a heuristic construction on finite measure spaces before considering a more formal, but also more abstract, construction that applies to any measure space. Next we'll investigate how the specification of expectation values can be used to implicitly define measures without having to explicitly define subset allocations and some useful applications. Finally we'll consider expectation values that are informed by common ambient space structures and then conclude with a discussion of a few exceptional measures whose expectation values can be computed algorithmically.

1 Expectation on Finite Measure Spaces

To start our discussion of expectation as simply as possible let's begin by considering a finite measure space compromised of the finite set X, a measure defined by the mass function

$$\mu : X \to [0, \infty],$$

and a real-valued function $f : X \to \mathbb{R}$.

The allocations defined by the mass function weights the elements of X relative to each other, emphasizing some while suppressing others. At the same time the function f associates those elements with a numerical output. We can then weight the numerical outputs by combining the weights of the inputs $\mu(x)$ and the individual output values $f(x)$,

$$\mu(x)\, f(x).$$

Adding all of these weighted outputs together gives a single number that is sensitive to the interplay between μ and f,

$$\sum_{x \in X} \mu(x) \cdot f(x).$$

The summary emphasizes not only large output values but also outputs from highly-weighted inputs. More formally this procedure defines the **expectation value** of f with respect to μ,

$$\mathbb{E}_{\mu}[f] \equiv \sum_{x \in X} \mu(x) \cdot f(x).$$

We use square brackets instead of round brackets to hint that the mapping doesn't take points as input but rather entire functions.

An interesting side effect of this construction is that expectation values are linear: given two real-valued functions $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ and two real constants $\alpha, \beta \in \mathbb{R}$ we have

$$\begin{aligned}
\mathbb{E}_{\mu}[\alpha \cdot f + \beta \cdot g] &= \sum_{x \in X} \mu(x) \cdot [\alpha \cdot f(x) + \beta \cdot g(x)] \\
&= \sum_{x \in X} \mu(x) \cdot (\alpha \cdot f(x) + \beta \cdot g(x)) \\
&= \alpha \cdot \sum_{x \in X} \mu(x) \cdot f(x) + \beta \cdot \sum_{x \in X} \mu(x) \cdot g(x) \\
&= \alpha \cdot \mathbb{E}_{\mu}[f] + \beta \cdot \mathbb{E}_{\mu}[g].
\end{aligned}$$

We will exploit this linearity property endlessly when we start appying expectation values in practice.

The expectation $\mathbb{E}_{\mu}[f]$ is sensitive to the behavior of f, but only in the context of μ. By considering multiple *test* measures, however, we can use this operation to more fully probe the behavior of a fixed function f. Expectations of f with respect to test measures that emphasizes certain input elements will be more sensitive to the corresponding output elements, probing different aspects of f. More intuitively we can interpret each test measure μ_j as encoding a question about f and the corresponding expectation value $\mathbb{E}_{\mu_j}[f]$ as encoding the answer (Figure 1a).

For example consider a *singular* probability mass function that concentrates entirely on a single element,

$$\delta_{x'}(x) = \begin{cases} 1, & x = x' \\ 0, & x \neq x' \end{cases}.$$

The expectation value of any real-valued function f with respect to δ_{x_i} is given by

$$\begin{aligned}
\mathbb{E}_{\delta_{x'}}[f] &= \sum_{x \in X} \delta_{x'}(x) \cdot f(x) \\
&= \delta_{x'}(x') \cdot f(x') + \sum_{x \neq x'} \delta_{x'}(x) \cdot f(x) \\
&= 1 \cdot f(x') + \sum_{x \neq x'} 0 \cdot f(x) \\
&= f(x_i).
\end{aligned}$$

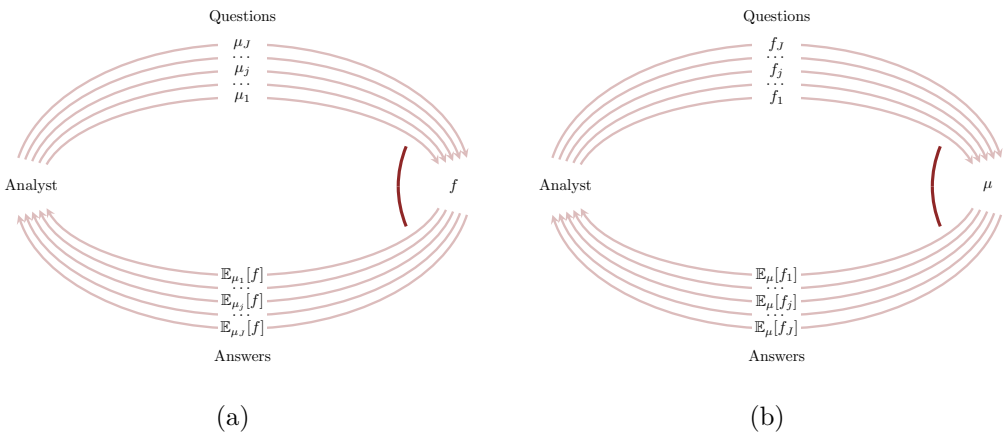

Figure 1: Expectations values probe the interaction between a measure and a real-valued function. (a) If we fix the function f then expectation values with respect to multiple test measures are sensitive to different features of f. We can interpret each test measure as a question about f with the corresponding expectation value providing an answer. (b) Similarly expectation values of multiple test functions with respect to a fixed measure μ are sensitive to different features of μ. Again we can interpret each test function as a question about μ with the corresponding expectation values encoding an answer.

In other words expectation values of functions with respect to $\delta_{x'}$ allow us to probe individual output values $f(x')$.

Similarly consider a uniform probability mass function where each element is allocated the same probability,

$$\pi(x) = \frac{1}{I}.$$

The corresponding expectation value captures the **average** of the function output values,

$$\begin{aligned}
\mathbb{E}_{\pi}[f] &= \sum_{x \in X} \pi(x) \cdot f(x) \\
&= \sum_{x \in X} \frac{1}{I} \cdot f(x) \\
&= \frac{1}{I} \sum_{x \in X} f(x).
\end{aligned}$$

When we use non-uniform measures this expectation operation generalizes averages to more general summaries.

At the same time we can use different test functions to probe different features of a fixed measure μ. Expectations of test functions with larger outputs for some inputs will be more

sensitive to the measures μ allocated to those inputs. Again we can interpret each test function f_j as encoding a different question about μ with the corresponding expectation value $\mathbb{E}_\mu[f_j]$ encoding the answer (Figure 1b).

For example for any subset $\mathsf{x} \subset X$ we can construct an **indicator function** that returns 1 if the input is contained in x and zero otherwise,

$$\mathbb{I}_{\mathsf{x}}(x) = \begin{cases} 1, & x \in \mathsf{x} \\ 0, & x \notin \mathsf{x} \end{cases}.$$

The expectation of the indicator function $\mathbb{I}_{\mathsf{x}}$, however, is just the measure allocated to x,

$$\begin{aligned}
\mathbb{E}_\mu[\mathbb{I}_{\mathsf{x}}] &= \sum_{x \in X} \mu(x) \cdot \mathbb{I}_{\mathsf{x}}(x) \\
&= \sum_{x \in \mathsf{x}} \mu(x) \cdot \mathbb{I}_{\mathsf{x}}(x) + \sum_{x \notin \mathsf{x}} \mu(x) \cdot \mathbb{I}_{\mathsf{x}}(x) \\
&= \sum_{x \in \mathsf{x}} \mu(x) \cdot 1 + \sum_{x \notin \mathsf{x}} \mu(x) \cdot 0 \\
&= \sum_{x \in \mathsf{x}} \mu(x) \\
&= \mu(\mathsf{x}).
\end{aligned}$$

Expectation values of various indicator functions allow us to directly probe the various subset allocations.

2 Expectation on General Measure Spaces

Unfortunately the straightforward construction of expectation values on finite spaces doesn't generalize to general measure spaces. In particular on uncountable spaces, where element-wise allocations $\mu(\{x\})$ do not completely characterize a measure, the weighted output values $\mu(\{x\}) \cdot f(x)$ do not completely characterize the interaction between a measure and a real-valued function.

In order to generalize expectation values to arbitrary measure spaces we have to appeal to a more sophisticated construction with some subtle, but important, consequences.

2.1 Expectation of Simple Functions

We'll build up to general expectation values by considering increasingly sophisticated classes of functions that are still nice enough for their expectation values to be unambiguous on any measurable space.

For example consider a constant function that always returns the same output for any input,

$$c(x) \mapsto c_0 \in \mathbb{R}$$

for all $x \in X$. The only possible expectation value for a constant function is that common output, in which case we should have

$$\mathbb{E}_\mu[c] = c_0$$

for *any* measure μ.

Increasing the complexity slightly let's next consider indicator functions which vanish outside of a given measurable subset (Figure 2)

$$\mathbb{I}_{\mathsf{x}}(x) = \begin{cases} 1, & x \in \mathsf{x} \\ 0, & x \notin \mathsf{x} \end{cases}.$$

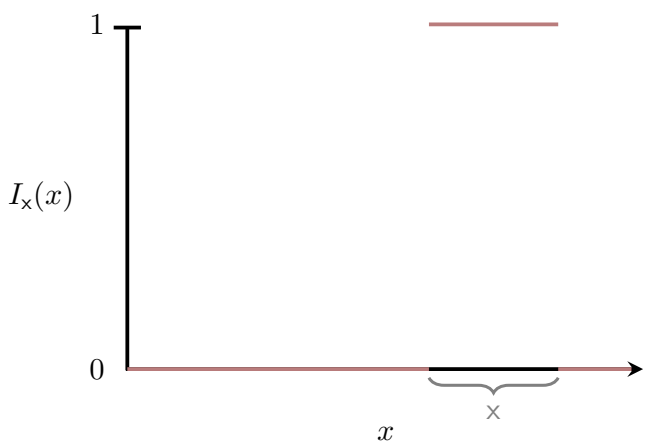

Figure 2: An indicator function corresponding to a measurable subset $\mathsf{x} \in \mathcal{X}$ vanishes for all inputs that are not contained in x. Here x is an interval subset over the real line.

In order to generalize the behavior on finite measure spaces that we encountered in Section One the expectation value of any indicator function should be equal to the measure allocated to that subset,

$$\mathbb{E}_\mu[\mathbb{I}_{\mathsf{x}}] = \mu(\mathsf{x}),$$

for any measure μ.

We can manufacture even more complex functional behavior still by overlaying multiple indicator functions on top of each other. A **simple function** is given by the sum of scaled indicator functions (Figure 3),

$$s(x) = \sum_j \phi_j \cdot I_{\mathsf{x}_j}(x),$$

where

$$\{\mathsf{x}_1, \dots, \mathsf{x}_j, \dots\} \in \mathcal{X}$$

is any sequence of measurable subsets and

$$\{\phi_1, \dots, \phi_j, \dots\} \in \mathbb{R}$$

is any sequence of real numbers. By incorporating countably many indicator functions we can engineer quite sophisticated functional behavior.

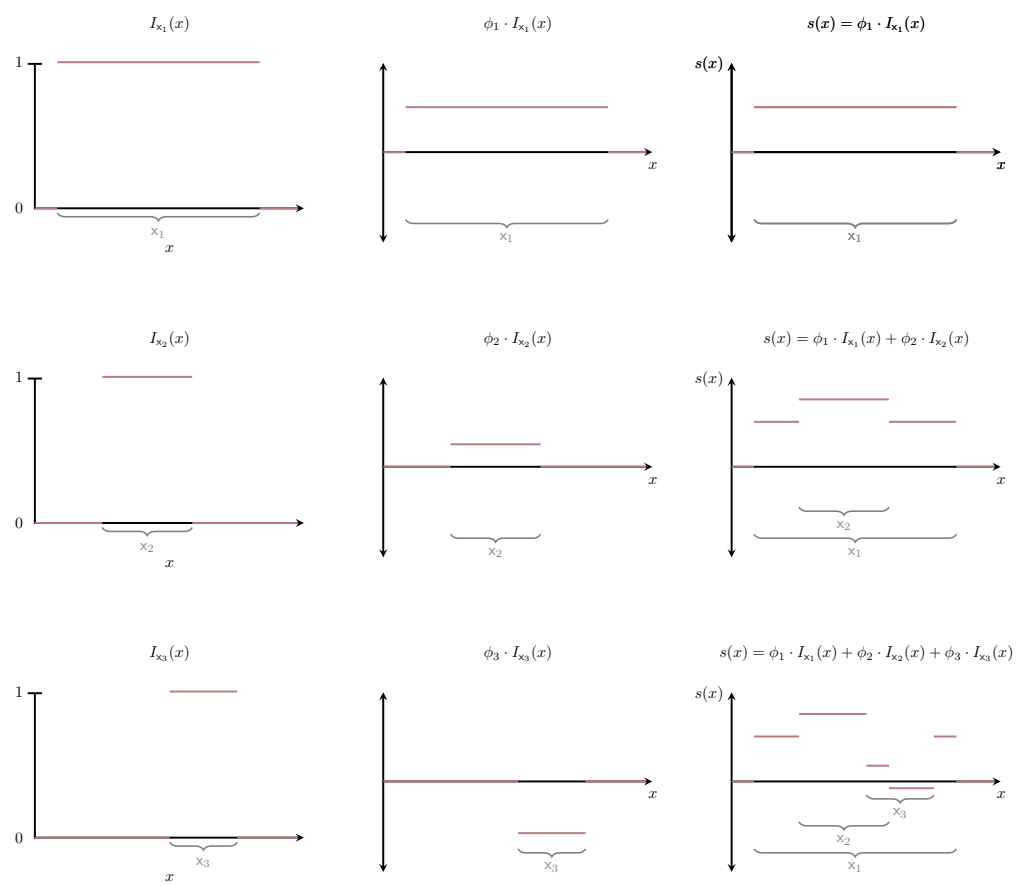

Figure 3: Simple functions are constructed from linear combinations of indicator functions. Incorporating more indicator functions yields more sophisticated functional behavior.

If we assume that expectation is a linear operation on any measure space then we can imme-

diately compute the expectation of any simple function,

$$\begin{aligned}
\mathbb{E}_\mu[s] &= \mathbb{E}_\mu \left[\sum_j \phi_j \cdot I_{\mathsf{x}_j} \right] \\
&= \sum_j \phi_j \cdot \mathbb{E}_\mu[I_{\mathsf{x}_j}] \\
&= \sum_j \phi_j \cdot \mu(\mathsf{x}_j).
\end{aligned}$$

2.2 Lebesgue Integration

Most functions whose expectation values are of interest in practical analysis are not simple functions, but they can often be *well-approximated* by simple functions. As we add more and more indicator functions we can construct simple functions that approximate non-simple functions better and better (Figure 4).

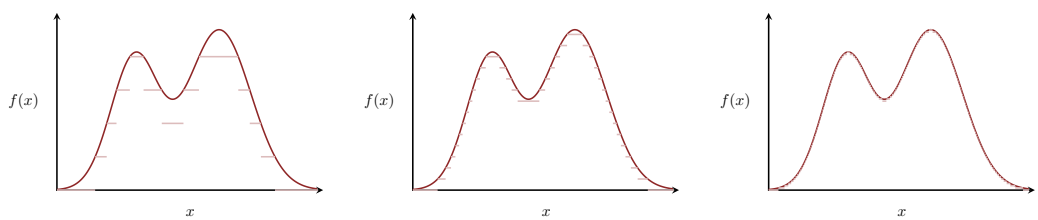

Figure 4: As we add more indicator functions simple functions become more flexible and are able to better approximate the behavior of non-simple functions. Certain non-negative functions can be *exactly* recovered from sufficiently flexible simple functions.

Some functions can even be *exactly* recovered from sufficiently flexible simple functions. A real-valued function $f : X \to \mathbb{R}$ is **measurable** with respect to the σ-algebra $\mathcal{X}$, or **$\mathcal{X}$-measurable**, if every half interval of outputs

$$(-\infty, x] \subset \mathbb{R}$$

pulls back to a measurable subset on $(X, \mathcal{X})$,

$$f^*((-\infty, x]) = \{x \in X \mid f(x) \in (\infty, x]\} \in \mathcal{X}.$$

We'll come back to the topic of measurable functions in much more detail in Chapter Seven, but for now our main concern will be to avoiding confusing measurable *subsets* on $(X, \mathcal{X})$ and measurable *functions* from $(X, \mathcal{X})$ to $\mathbb{R}$.

Measurable functions with non-negative outputs,

$$f(x) \geq 0$$

for all $x \in X$, are particularly special. Any non-negative, measurable function can always be perfectly recovered as a certain limit of increasingly complicated simple functions,

$$f(x) = \sum_{j=1}^{\infty} \phi_j \cdot I_{\mathsf{x}_j}(x).$$

We can then *define* the expectation value of a non-negative, measurable function as the expectation value of the corresponding simple function decomposition,

$$\begin{aligned} \mathbb{E}_\mu[f] &\equiv \mathbb{E}_\mu \left[\sum_{j=1}^{\infty} \phi_j \cdot I_{\mathsf{x}_j} \right] \\ &= \sum_{j=1}^{\infty} \phi_j \cdot \mathbb{E}_\mu \left[I_{\mathsf{x}_j} \right] \\ &= \sum_{j=1}^{\infty} \phi_j \cdot \mu(\mathsf{x}_j). \end{aligned}$$

In general a non-negative, measurable function can be constructed from multiple simple functions, but the expectation values derived from any of them will all be the same. Consequently there's no worry for ambiguous of otherwise inconsistent answers, and expectation values for non-negative, measurable function are completely well-behaved. This procedure for defining expectation values through simple functions representations is known as **Lebesgue integration**.

We've come a long way, but non-negative functions are still somewhat exceptional amongst all of the functions that might come up in a given analysis. To define expectation values for measurable functions that aren't necessarily positive we just have to decompose the functions by the sign of their outputs (Figure 5),

$$f(x) = f^+(x) - f^-(x),$$

where

$$f^+(x) = \begin{cases} f(x), & f(x) \geq 0 \\ 0, & f(x) < 0 \end{cases}$$

and

$$f^-(x) = \begin{cases} -f(x), & f(x) < 0 \\ 0, & f(x) \geq 0 \end{cases}.$$

Because f^+ and f^- are both non-negative we can construct their expectation values $\mathbb{E}_\mu[f^+]$ and $\mathbb{E}_\mu[f^-]$ as above. Provided that the expectation values are not both infinite we can then define the expectation value of f by taking advantage of linearity,

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[f^+ - f^-] = \mathbb{E}_\mu[f^+] - \mathbb{E}_\mu[f^-].$$

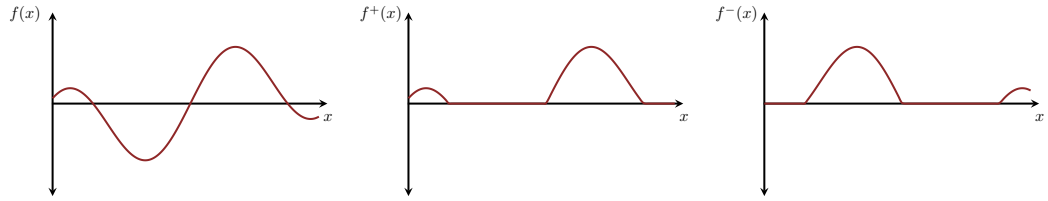

Figure 5: Every real-valued function $f : X \to \mathbb{R}$ function can be decomposed by the sign of its output values, resulting in the two positive functions $f^+ : X \to \mathbb{R}^+$ and $f^- : X \to \mathbb{R}^+$.

One way to ensure that this difference is well-defined is to require that

$$\mathbb{E}_\mu[|f|] = \mathbb{E}_\mu[f^+ + f^-] = \mathbb{E}_\mu[f^+] + \mathbb{E}_\mu[f^-]$$

is finite. Measurable functions $f : X \to \mathbb{R}$ with

$$\mathbb{E}_\mu[|f|] < \infty$$

are said to be **Lebesgue integrable** with respect to μ, or just **μ-integrable** for short.

Nearly every real-valued function that we will encounter in practical applications will be measurable. Consequently taking this technical assumption for granted is largely safe. Many real-valued functions will also be integrable with respect to typical measures, especially when we restrict attention to probability distributions, but there are enough exceptions that we have to be careful to explicitly validate integrability in practice.

To streamline the writing I will take advantage of some more casual vocabulary for the remainder of this book. I will refer to any real-valued function $f : X \to \mathbb{R}$ that is measurable with respect to the ambient σ-algebra and integrable with respect to any relevant measures simply as an **expectand**. This is a bit of a terminological analogy with calculus: an *integrand* is a sufficiently well-behaved function whose integral is well-defined while an *expectand* is a sufficiently well-behaved function whose expectation value is well-defined.

2.3 Equivalent Expectands

One subtle but important consequence of this general definition of expectation is that many expectands will yield the same expectation values even when their individual outputs are not all equal!

To see why let's consider a simple function $s : X \to \mathbb{R}$ that's build up from arbitrarily many indicator functions,

$$s(x) = \sum_j \phi_j \cdot I_{\mathsf{x}_j}(x).$$

Adding another indicator function with respect to the measurable subset x' gives another simple function,

$$s'(x) = s(x) + \phi' \cdot I_{\mathsf{x}'}(x).$$

The expectation values of these two simple functions are then related to each other by

$$\begin{aligned}
\mathbb{E}_\mu[s'] &= \mathbb{E}_\mu[s + \phi' \cdot I_{\mathsf{x}'}] \\
&= \mathbb{E}_\mu[s] + \phi' \cdot \mathbb{E}_\mu[I_{\mathsf{x}'}] \\
&= \mathbb{E}_\mu[s] + \phi' \cdot \mu(\mathsf{x}').
\end{aligned}$$

When $\phi' \neq 0$ then the outputs of s and s' will differ for all of the inputs in x'; so long as x' is not the empty set then the function outputs will differ for at least some inputs. On the other hand the corresponding expectation values will differ only if

$$\mu(\mathsf{x}') > 0!$$

In other words if x' is a μ-null subset then s and s' will share the exact same μ-expectation values.

More generally any two expectands $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ will share the same μ-expectation value if the subset of input points where their outputs differ,

$$\mathsf{x}_\delta = \{x \in X \mid f(x) \neq g(x)\},$$

is a subset of μ-null subset,

$$\mathsf{x}_\delta \subseteq \mathsf{x} \in \mathcal{X}$$

with

$$\mu(\mathsf{x}) = 0.$$

Ultimately modifying expectands on sets of measure zero does not affect their expectation values.

If the subset of deviant inputs is contained within a μ-null subset f and g are said to be equal **almost everywhere** with respect to μ. When working with probability distributions instead of measures the term **almost surely** equal is used instead. A bit more colloquially we can say that the two expectands are equal **up to subsets of measure zero** or equal **up to null subsets**. I will use this latter terminology for the remainder of the book.

Intuitively the null subsets of a measure can "wash out" some of the finer structure of expectands. For example on a real line any countable collection of points is allocated zero Lebesgue measure. Consequently expectation with respect to the Lebesgue measure will disregard any "point defects" in the expectands (Figure 6).

Applications of measure theory can't distinguish between expectands that are equal up to sets of measure zero. If we want to avoid this ambiguity then we have to impose structural constraints on the equivalent expectands to isolate a single, unique expectand. For example

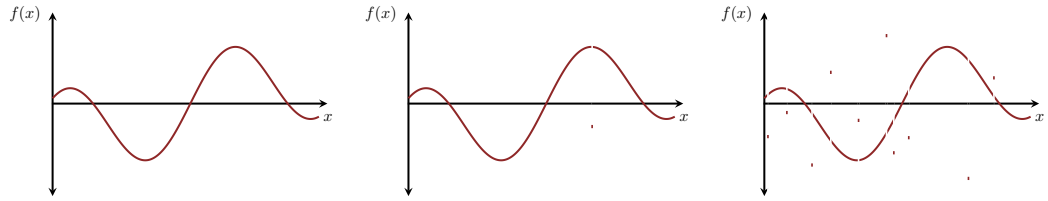

Figure 6: Because any countable collection of points is allocated zero Lebesgue measure any expectands whose outputs differ only at a countable number of inputs will yield the same expectation values. In other words from the perspective of the Lebesgue measure these functions are equivalent.

we can modifying a continuous expectand on input subsets of measure zero without changing the expectation value, but those modifications will also introduce discontinuities. Amongst all of the equivalent expectands only one will be continuous. Consequently even though general expectands are not unique continuous expectands are.

Equality up to sets of measure zero is mostly a technical concern, but there are few exceptional circumstances where it will be relevant in practice. I will clearly point these circumstances out as we go along.

2.4 Alternative Expectation Notations

One of the limitations of the standard expectation value notation, $\mathbb{E}_{\mu}[f]$, is that it doesn't denote the ambient space. When working on a single space this isn't too much of an issue, but it can cause confusion when we start working with multiple spaces at the same time.

A more expressive notation like

$$\mathbb{E}_{(X,\mathcal{X},\mu)}[f]\mathbb{E}_{(Y,\mathcal{Y},\nu)}[g]$$

is much more explicit but also much more cumbersome. Mathematicians have developed a variety of shorthand notations that offer different compromises between clarity and compactness.

For example some references denote expectation values as

$$\mathbb{E}_{\mu}[f] = \int_X \mu\, f,$$

where the subscript of the integral sign allows us to specify the ambient space and a σ-algebra is taken for granted. When using this notation, however, we have to be careful to not confuse $\int$ with the standard integral from calculus. We'll discuss the subtle relationship between expectation values and integrals in more detail in Section 5.2.

We can also use variables to denote the ambient space. Taking $x \in X$ to be a variable that takes values in X some references denote expectation values as

$$\mathbb{E}_\mu[f] = \int \mu(\mathrm{d}x)\, f(x),$$

or

$$\mathbb{E}_\mu[f] = \int \mathrm{d}\mu(x)\, f(x).$$

The placement of the measure and the expectand is conventional; some references prefer instead

$$\mathbb{E}_\mu[f] = \int f(x)\, \mu(\mathrm{d}x),$$

and

$$\mathbb{E}_\mu[f] = \int f(x)\, \mathrm{d}\mu(x).$$

Again when using these particular notations we have to be careful to avoid confusing them with integrals.

For this book I will use $\mathbb{E}_\mu[f]$ most often, but when it becomes convenient I'll also use the notation

$$\mathbb{E}_\mu[f] = \int \mu(\mathrm{d}x)\, f(x).$$

3 Specifying Measures With Expectation Values

To this point we have derived expectation values as a consequence of measurable subset allocations. Expectation values, however, can also be used to define measures directly, with subset allocations derived from appropriate expectation values.

While a bit more abstract than our initial approach this perspective does have its benefits.

3.1 Functional Perspective of Measures

Expectation values map real-valued functions into real numbers. If we denote the space of all functions from X to $\mathbb{R}$ as $C(X)$ then we might be tempted to write this mapping as

$$\begin{aligned} \mathbb{E}_\mu : C(X) &\to \mathbb{R} \\ f \quad &\mapsto \mathbb{E}_\mu[f]. \end{aligned}$$

Unfortunately this isn't technically correct because not every real-valued function has a well-defined expectation value. In other words $C(X)$ is too large of an input space.

To remedy that we can define

$$L(X, \mathcal{X}, \mu) \subset C(X)$$

as the subset of real-valued functions from X to $\mathbb{R}$ that are measurable with respect to $\mathcal{X}$ and then integrable with respect to μ. Using the terminology introduced in the previous section, $L(X, \mathcal{X}, mu)$ is the space of expectands.

With this notation expectation can be interpreted as a map from expectands to real numbers,

$$\begin{aligned} \mathbb{E}_\mu : L(X, \mathcal{X}, \mu) &\to \mathbb{R} \\ f \qquad\qquad &\mapsto \mathbb{E}_\mu[f]. \end{aligned}$$

In fact expectation is the only *linear* map of this form.

Because $L(X, \mathcal{X}, \mu)$ contains all of the indicator functions this functional relationship between $L(X, \mathcal{X}, \mu)$ and $\mathbb{R}$ determines the allocations to every measurable subset, and hence full determines the measure μ. At the same time $L(X, \mathcal{X}, \mu)$ also contains many expectands that are not indicator functions, and hence quite a bit of redundant information about μ.

Sufficiently nice measures can be completely characterized by their expectation action on subsets of $L(X, \mathcal{X}, \mu)$ that do not contain any indicator functions at all! In theory the expectation values of other expectands, including indicator functions to recover subset allocations, can then be derived from these initial expectands. These sparser characterizations are particularly useful for analyzing certain theoretical properties of measures with the tools of **functional analysis**.

The expectation perspective also has its benefits for applied practice. For example once we've built a probability distribution relevant to an application we will use expectation values to extract meaningful information. **Probabilistic computational algorithms** automate this process, mapping expectands to expectation values exactly, or more realistically, approximately.

Interpreting measures as expectation value generators helps us understand not only what operations we need to carry out to realize an applied analysis but also how well our algorithmic tools actually implement those operations. We will spend a good bit of time discussing these issues in later chapters.

3.2 Scaling Measures

Measures become much more flexible tools when we can readily modify their behavior, enhancing the measure at some points while suppressing it at others. The functional perspective of measures is particularly convenient for implicitly defining these modifications that would be at best awkward to specify directly through subset allocations.

For example let's say that we want to *globally* scale the subset allocations defined by μ with a constant $\alpha \in \mathbb{R}^+$. The scaled measure is straightforward to define by modifying the individual subset allocations,

$$(\alpha \cdot \mu)(\mathsf{x}) \equiv \alpha \cdot \mu(\mathsf{x})$$

for all measurable subsets $\mathsf{x} \in \mathcal{X}$.

These scaled allocations then imply that expectation values of simple functions with respect to $\alpha \cdot \mu$ can be recovered as expectation values of *scaled* expectands with respect to μ,

$$\begin{aligned}
\mathbb{E}_{\alpha \cdot \mu}[s] &= \mathbb{E}_{\alpha \cdot \mu}\left[\sum_j \phi_j \cdot I_{\mathsf{x}_j}\right] \\
&= \sum_j \phi_j \cdot \mathbb{E}_{\alpha \cdot \mu}[I_{\mathsf{x}_j}] \\
&= \sum_j \phi_j \cdot (\alpha \cdot \mu)(\mathsf{x}_j) \\
&= \sum_j \phi_j \cdot \alpha \cdot \mu(\mathsf{x}_j) \\
&= \alpha \cdot \sum_j \phi_j \cdot \mu(\mathsf{x}_j) \\
&= \alpha \cdot \sum_j \phi_j \cdot \mathbb{E}_{\mu}[I_{\mathsf{x}_j}] \\
&= \mathbb{E}_{\mu}\left[\alpha \cdot \sum_j \phi_j \cdot I_{\mathsf{x}_j}\right] \\
&= \mathbb{E}_{\mu}[\alpha \cdot s].
\end{aligned}$$

Because general expectation values are derived from the expectation values of simple functions we will then have

$$\mathbb{E}_{\alpha \cdot \mu}[f] = \mathbb{E}_{\mu}[\alpha \cdot f]$$

for *every* expectand $f : X \to \mathbb{R}$. In other words these modified expectation values fully define the scaled measure $\alpha \cdot \mu$ just as well as the modified subset allocations.

To complicate matters we might then ask how we can *locally* scale a measure by some positive, $\mathcal{X}$-measurable, real-valued function $g : X \to \mathbb{R}^+$. Because g varies across non-atomic subsets it is no longer clear how we can consistently modify all of the initial subset allocations to be larger when g is larger and smaller when g is smaller.

The functional construction, however, immediately generalizes. We can *define* a scaled measure $g \cdot \mu$ as the unique measure with the expectation values

$$\mathbb{E}_{g \cdot \mu}[f] \equiv \mathbb{E}_{\mu}[g \cdot f]$$

for every expectand $f : X \to \mathbb{R}$ with

$$\mathbb{E}_{\mu}[|g \cdot f|] < \infty.$$

This expectation value definition can then be used to calculate the subtle, but necessary, modifications to the subset allocations,

$$\begin{aligned}(g \cdot \mu)(\mathsf{x}) &= \mathbb{E}_{g \cdot \mu}[I_{\mathsf{x}}] \\ &= \mathbb{E}_{\mu}[g \cdot I_{\mathsf{x}}].\end{aligned}$$

In particular the modified subset allocations are no longer given by simple scalings of the initial subset allocations!

This flexible construction can then be applied in a variety of useful ways. For example scaling a measure μ by the indicator function of a measurable subset x',

$$\mathbb{E}_{I_{\mathsf{x}'} \cdot \mu}[f] \equiv \mathbb{E}_{\mu}[I_{\mathsf{x}'} \cdot f],$$

consistently zeroes out all measure outside of x', restricting μ to that subset. If X is an ordered space and x' is an interval subset then this restriction is also known as **truncation**.

3.3 Scaling Probability Distributions

Scaling probability distributions is not quite as straightforward because we have to maintain the proper normalization. Naively scaling a probability distribution π with a positive, $\mathcal{X}$-measurable, real-valued function $g : X \to \mathbb{R}^{+}$ results in a total measure

$$\begin{aligned}(g \cdot \pi)(X) &= \mathbb{E}_{g \cdot \pi}[I_X] \\ &= \mathbb{E}_{g \cdot \pi}[1] \\ &= \mathbb{E}_{\pi}[g]\end{aligned}$$

which is not, in general, equal to 1. In other words scaling a probability distribution results not in another probability distribution but rather a generic measure.

If we want transform one probability distribution into another then we need to correct for the modified normalization, defining

$$\mathbb{E}_{g * \pi}[f] \equiv \frac{\mathbb{E}_{\pi}[g \cdot f]}{\mathbb{E}_{\pi}[g]}$$

for every expectand $f : X \to \mathbb{R}$ with

$$\mathbb{E}_{\pi}[|g \cdot f|] < \infty.$$

In this case the modfied subset allocations become

$$\begin{aligned}(g * \pi)(\mathsf{x}) &= \frac{\mathbb{E}_{g \cdot \pi}[I_{\mathsf{x}}]}{\mathbb{E}_{\pi}[g]} \\ &= \frac{\mathbb{E}_{\pi}[g \cdot I_{\mathsf{x}}]}{\mathbb{E}_{\pi}[g]}.\end{aligned}$$

Specifically we will always have

$$\begin{aligned}(g * \pi)(X) &= \frac{\mathbb{E}_{\pi}[g \cdot I_X]}{\mathbb{E}_{\pi}[g]} \\ &= \frac{\mathbb{E}_{\pi}[g]}{\mathbb{E}_{\pi}[g]} \\ &= 1\end{aligned}$$

as necessary.

For example scaling with an indicator function restricts a probability distribution to the corresponding subset and reduces the total probability to the probability initially allocated to that subset,

$$\begin{aligned}(I_{\mathsf{x}'} \cdot \pi)(X) &= \mathbb{E}_{I_{\mathsf{x}'} \cdot \pi}[I_X] \\ &= \mathbb{E}_{I_{\mathsf{x}'} \cdot \pi}[1] \\ &= \mathbb{E}_{\pi}[I_{\mathsf{x}'}] \\ &= \pi(\mathsf{x}').\end{aligned}$$

Scaling and then normalizing, however, corrects the proportional subset allocations to this restriction,

$$\begin{aligned}(I_{\mathsf{x}'} * \pi)(\mathsf{x}) &= \frac{\mathbb{E}_{I_{\mathsf{x}'} * \pi}[I_{\mathsf{x}}]}{\mathbb{E}_{\pi}[I_{\mathsf{x}'}]} \\ &= \frac{\mathbb{E}_{\pi}[I_{\mathsf{x}'} \cdot I_{\mathsf{x}}]}{\mathbb{E}_{\pi}[I_{\mathsf{x}'}]} \\ &= \frac{\mathbb{E}_{\pi}[I_{\mathsf{x}' \cap \mathsf{x}}]}{\mathbb{E}_{\pi}[I_{\mathsf{x}'}]} \\ &= \frac{\pi(\mathsf{x}' \cap \mathsf{x})}{\pi(\mathsf{x}')}.\end{aligned}$$

In particular

$$(I_{\mathsf{x}'} * \pi)(X) = \frac{\pi(\mathsf{x}' \cap X)}{\pi(\mathsf{x}')} = \frac{\pi(\mathsf{x}')}{\pi(\mathsf{x}')} = 1.$$

4 Structure-Informed Expectation Values

Every ambient space admits infinitely many real-valued functions, and hence endless ways to interrogate a given measure through expectation values. Some expectands, however, are uniquely compatible with the structure of the space itself and their expectation values extract particularly interpretable information. In this section we'll review some of the most common of these structure-informed expectands.

4.1 Moments and Cumulants

Some spaces are inherently related to a real line. The precise relationship between elements of a space and elements of a real line defines a distinguished real-valued function ,and hence a distinguished expectand. We can even build off of this initial expectand to construct an entirely family of useful expectands.

4.1.1 Embeddings

In order for an ambient space X to be compatible with the real line it needs to share a metric structure. We say that we can **embed** a metric space X into a real line if we can construct an isometric injection $\iota : X \to \mathbb{R}$, i.e. a function that maps each element of X to a distinct output while also preserving distances,

$$d_X(x_1, x_2) = d_{\mathbb{R}}(\iota(x_1), \iota(x_2)) = |\iota(x_2) - \iota(x_1)|.$$

Embedding maps are often denoted with a hooked arrow instead of the typical flat arrow,

$$\iota : X \hookrightarrow \mathbb{R},$$

to communicate that some structure is being preserved.

For example if X is itself a real line then the identify map defines a natural embedding $\iota : \mathbb{R} \hookrightarrow \mathbb{R}$ (Figure 7a). Similarly we can embed subsets of the real line, such as intervals $\iota : [x_1, x_2] \hookrightarrow \mathbb{R}$ (Figure 7b) or even integers $\iota : \mathbb{Z} \hookrightarrow \mathbb{R}$ (Figure 7c).

The existence of an embedding map can be interpreted in a few different ways. On one hand it implies that X is isomorphic to some subset of a real line, if not an entire real line, which allows us to interpret X as a subset of a real line. Alternatively we can think of an embedding map as assigning to each element $x \in X$ a numerical position that we can use to characterize geometric behavior. Both interpretations are useful but in this section we will lean heavily on this latter perspective.

When an embedding map is measurable with respect to the ambient σ-algebra and integrable with respect to the ambient measure it defines an expectand. Most embedding maps are measurable but integrability is less dependable, and failures of integrability are important in practice.

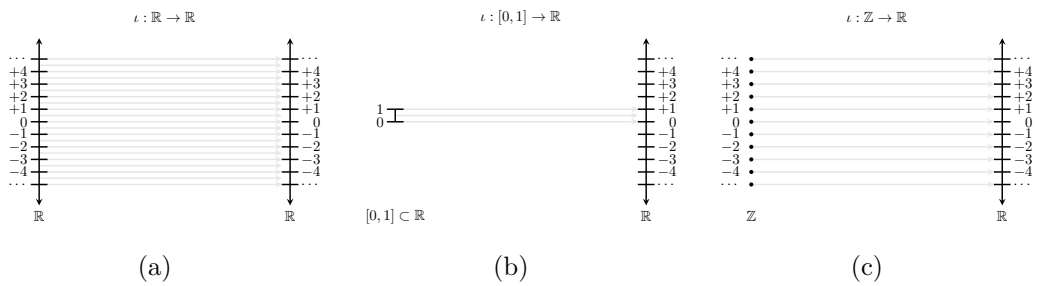

Figure 7: Many spaces naturally embed into a real line, including (a) that real line, (b) intervals of that real line, and (c) integers.

4.1.2 The Mean

If an embedding function is an expectand than we can evaluate its expectation value, $\mathbb{E}_\mu[\iota]$. The ultimately utility of this expectation value, however, depends on what information about the ambient measure it extracts.

Interpreting $\mathbb{E}_\mu[\iota]$ is straightforward when X is finite, $\mathcal{X}$ is the full power set, and we can represent any measure with a mass function. In this case we can explicitly compute the $\mathbb{E}_\mu[\iota]$ as a weighted sum of positions,

$$\mathbb{E}_\mu[\iota] = \sum_{x \in X} \mu(x)\, \iota(x).$$

The more measure that is allocated to an element the more strongly the expectation value is pulled towards the position of that element. In other words $\mathbb{E}_\mu[\iota]$ is one way to quantify the position around which the measure μ concentrates, defining a notion of *centrality* for the measure μ.

This interpretation does generalize to arbitrary spaces, although the formal motivation is a bit more subtle because we can no longer interpret expectation values as simple weighted sums. Instead consider a baseline position $r_0 \in \mathbb{R}$ and the squared distance function

$$\begin{aligned} d^2_{r_0} &: X \to \mathbb{R}^+ \\ x &\mapsto (\iota(x) - r_0)^2 \end{aligned}$$

which quantifies how far the position of any point in the ambient space is from that baseline position.

So long as $\iota : X \hookrightarrow \mathbb{R}$ is an embedding this squared distance function will be measurable and

will define a valid expectand. The expectation value

$$\begin{aligned}
\mathbb{E}_{\mu}\left[d_{r_0}^2\right] &= \mathbb{E}_{\mu}\left[(\iota - r_0)^2\right] \\
&= \mathbb{E}_{\mu}\left[\iota^2 - 2\, r_0\, \iota + r_0^2\right] \\
&= \mathbb{E}_{\mu}\left[\iota^2\right] - 2\, r_0 \cdot \mathbb{E}_{\mu}\left[\iota\right] + \mathbb{E}_{\mu}\left[r_0^2\right] \\
&= \mathbb{E}_{\mu}\left[\iota^2\right] - 2\, r_0 \cdot \mathbb{E}_{\mu}\left[\iota\right] + r_0^2
\end{aligned}$$

quantifies how diffusely the measure μ is allocated around r_0; the larger the expectation value the further r_0 is from the neighborhoods where μ concentrates its allocation.

Consequently the baseline position $r_0 \in \mathbb{R}$ with the *smallest* expected squared distance should be, in some sense, the position closest to where μ concentrates. Because we're working with continuous positions we can compute the baseline position that minimizes the expected squared distance using calculus methods even if X itself is not continuous.

In particular the minimum r_0^* is given by setting the derivative of the expectation value,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}r_0}\mathbb{E}_{\mu}\left[d_{r_0}^2\right] &= \frac{\mathrm{d}}{\mathrm{d}r_0}\left(\mathbb{E}_{\mu}\left[\iota^2\right] - 2\, r_0 \cdot \mathbb{E}_{\mu}\left[\iota\right] + r_0^2\right) \\
&= -2\, \mathbb{E}_{\mu}\left[\iota\right] + 2\, r_0,
\end{aligned}$$

to zero,

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}r_0}\mathbb{E}_{\mu}\left[d_{r_0}^2\right]\Bigg|_{r_0 = r_0^*} \\
0 &= -2\, \mathbb{E}_{\mu}\left[\iota\right] + 2\, r_0^* \\
2\, r_0^* &= 2\, \mathbb{E}_{\mu}\left[\iota\right] \\
r_0^* &= \mathbb{E}_{\mu}\left[\iota\right].
\end{aligned}$$

In other words the expectation value of the embedding function $\mathbb{E}_{\mu}[\iota]$ is exactly the position that minimizes the expected squared distance and, in that sense, is closest to the concentration of μ. Note that if X is not continuous, for example if $X = \mathbb{Z}$, then this central position might fall between the positions of the individual elements (Figure 8).

Because $\mathbb{E}_{\mu}\left[\iota\right]$ quantifies a sense of the centrality of the measure μ is referred to as the **mean** of μ, in reference to the "middle" of the measure. When space is at a premium I will use $\mathbb{M}_{\mu}$ to denote the mean, with the ambient space and embedding map all implicit.

4.1.3 Higher-Order Moments and Cumulants

We have not yet, however, exhausted the usefulness of an embedding map. For example the expected squared distance from the mean

$$\mathbb{E}_{\mu}\left[d_{\mathbb{M}_{\mu}}^2\right] = \mathbb{E}_{\mu}\left[(\iota - \mathbb{M}_{\mu})^2\right]$$

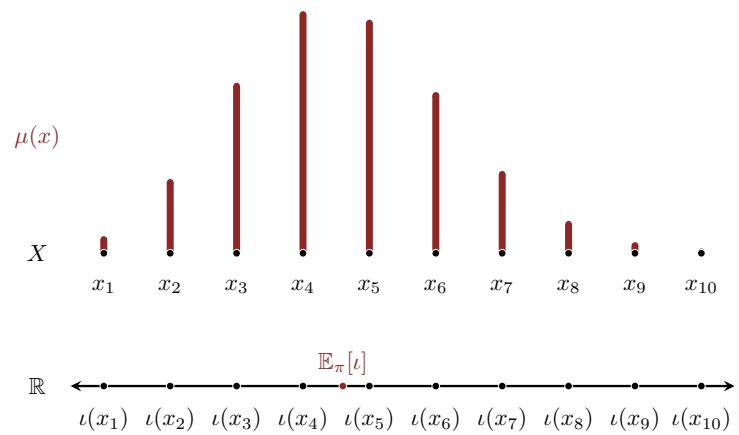

Figure 8: The expectation value of an embedding function, $\mathbb{E}_\mu[\iota]$, is a continuous value even when the ambient space is discrete. In these cases the centrality of a measure can fall between the individual elements.

quantifies how *strongly* μ concentrates around its centrality; the larger the expectation value the more diffuse the concentration is. This is known as the **variance** of μ.

Higher-powers extract even more information. For example the expectation value of the cubic expectand

$$\mathbb{E}_\mu \left[(\iota - \mathbb{M}_\mu)^3 \right]$$

characterizes how *symmetric* the concentration of μ is around its centrality.

The expectation values that we can construct from an embedding function can be systemized in various ways. For example the direct powers

$$\mathbb{M}_{\mu,k} = \mathbb{E}_\mu \left[\iota^k \right]$$

define the k-th order **moments** while the shifted powers

$$\mathbb{D}_{\mu,k} = \mathbb{E}_\mu \left[(\iota - \mathbb{M}_\mu)^k \right]$$

define the k-th order **central moments**. In some cases normalizing the central moments,

$$\mathbb{N}_{\mu,k} = \frac{\mathbb{E}_\mu \left[(\iota - \mathbb{M}_\mu)^k \right]}{(\mathbb{D}_{\mu,2})^{k/2}} = \frac{\mathbb{E}_\mu \left[(\iota - \mathbb{M}_\mu)^k \right]}{\left(\mathbb{E}_\mu \left[(\iota - \mathbb{M}_\mu)^2 \right] \right)^{k/2}},$$

to give k-th order **standardized central moments** is also useful.

While straightforward to construct, higher-order moments can be tricky to interpret. More useful information can often be isolated by carefully mixing a higher-order moment with lower-order moments, resulting in **cumulants**, $\mathbb{C}_{\mu,k}$. The general construction of cumulants is

complicated, with some very interesting but very elaborate connections to combinatorics, but in this book we'll focus on the first few cumulants. Conveniently the first-order cumulant is just the mean,

$$\mathbb{C}_{\mu,1} = \mathbb{M}_{\mu,1},$$

the second-order cumulant is just the variance,

$$\mathbb{C}_{\mu,2} = \mathbb{D}_{\mu,2},$$

and the third-order cumulant is just the third-order central moment,

$$\mathbb{C}_{\mu,3} = \mathbb{D}_{\mu,3}.$$

Beyond third-order the cumulants begin deviate away from the central moments.

4.1.4 Spaces Without Moments

Well-defined moments can be obstructed in a variety of different ways. For example if the embedding function and its powers are not integrable with respect to a measure then the corresponding expectation values will be ill-defined. The measurability of an embedding function is a more subtle, but still occasionally important, obstruction.

Exactly when an embedding function is measurable will depend on how the metric and σ-algebra on the ambient space interact with each other. For example consider a space X that is equipped with a metric and a metric topology. When we use the Borel σ-algebra derived from that topology then the metric and the σ-algebra will play nicely with each other.

In this case an embedding function will be measurable only if it is a homeomorphism from the ambient space into a real line. This, in turn, will be true only when the topology of the ambient space and the topology of the real line are compatible with each other. Consequently topological considerations can obstruct the existence of moments.

Recall that in Chapter Two we saw that the topologies of a circle, S^1, and a real line, $\mathbb{R}$, are fundamentally incompatible with each other. Because no real-valued function $f : S^1 \to \mathbb{R}$ is be a homeomorphism, no embedding is Borel measurable, and there is no well-defined notion of a mean or variance for any measure on the circle equipped with a Borel σ-algebra!

If you're not experienced with topology then this argument might appear to be overly technical and easy to dismiss, but the practical consequences are critical when working with spaces like circles, spheres, torii, and more. Many analyses have been sunk by attempts to estimate means on these spaces that don't actually exist!

Examples like this show why topology, as frustratingly abstract as it is often presented, is more important to applied practice than we might realize. Topological considerations can be extremely helpful in preventing us from succumbing to many common misconceptions.

4.2 Histograms

When the structure of the ambient space distinguishes certain subsets the corresponding indicator functions become natural expectands to consider. Conveniently the expectation values of indicator functions are also straightforward to interpret.

For example an ordering on the ambient space motivates interval subsets, such as the half-open interval subsets

$$(x_1, x_2] = \{x \in X \mid x_1 < x \leq x_2\}.$$

We can then use disjoint intervals to study the behavior of a measure by investigating how the measure allocations, or equivalently the expectation values of the corresponding indicator functions, vary across the ambient space.

More formally given the sequence of points

$$\{x_1, \dots, x_b, \dots, x_{B+1}\} \in X$$

we can partition a subset of the ambient space into a sequence of B disjoint half-open intervals,

$$\begin{aligned}
\mathsf{b}_1 &= (x_1, x_2] \\
\mathsf{b}_2 &= (x_2, x_3] \\
&\dots \\
\mathsf{b}_b &= (x_b, x_{b+1}] \\
&\dots \\
\mathsf{b}_B &= (x_B, x_{B+1}].
\end{aligned}$$

Evaluating the expectation value of the indicator function corresponding to each interval gives the allocated measure,

$$\mathbb{E}_\mu[I_{\mathsf{b}_b}] = \mu(\mathsf{b}_b).$$

Each of these measure allocations can then be neatly visualized as a rectangle, with the collection of measure allocations visualized as a sequence of adjacent rectangles (Figure 9). This visualization is referred to as a **histogram**, with the individual intervals denoted **bins**.

Histograms are incredibly useful for quickly communicating some of the key features of a measure (Figure 13). For example histograms allow us to differentiate between allocations that concentrate around a point, referred to as **unimodal** measures, or even allocations that concentrate around multiple points, referred to as **multimodal** measures. At the same time we can see *how* a measure concentrates around a point, for example whether the concentrations is symmetric or skewed towards smaller or larger values.

The smaller the bins the finer the features we can resolve but the more expectation values we have to compute in order to construct the histogram (Figure 11). In practice we have to

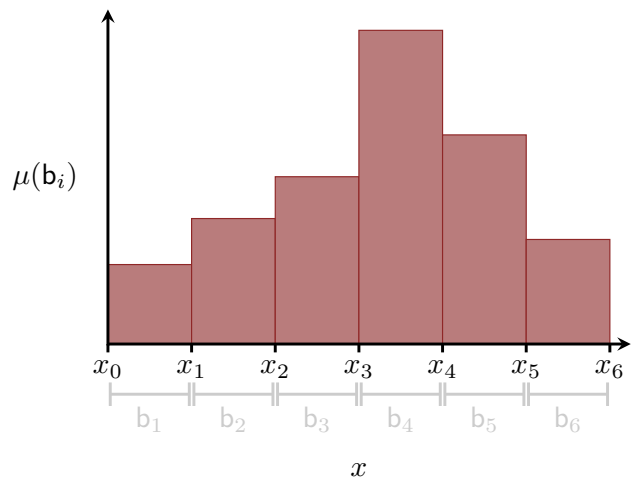

Figure 9: A histogram allows us to visualize the behavior of a measure over an ordered space. After partitioning a segment of the ambient space into disjoint intervals, or bins, the measure allocated to each bin is represented by a rectangle.

choose a binning that is suited to each measure of interest without being too expensive to implement.

The practical limitation of a finite number of bins also requires care in how we choose the boundaries of a histogram. Because a histogram censors any behavior below the first bin and above the last bin we need to choose the binning to span all of the behaviors of interest. For example if the ambient measure allocations decay towards smaller and larger values then we can set the bin boundaries where the allocations become negligible.

On countable spaces we can always tune the bins in a histogram to span only a single element. In this case the height of each bin reduces to $\mu(\{x\})$ and the resulting histogram then reduces to a visualization of the mass function.

4.3 Cumulative Distribution Functions

On an ordered space we can also use interval subsets to visualize how the total measure is allocated as we go from smaller values to larger values. More concretely consider the interval subsets consisting of all points smaller than or equal to a given point,

$$I_x = \{x' \in X \mid x' \leq x\}.$$

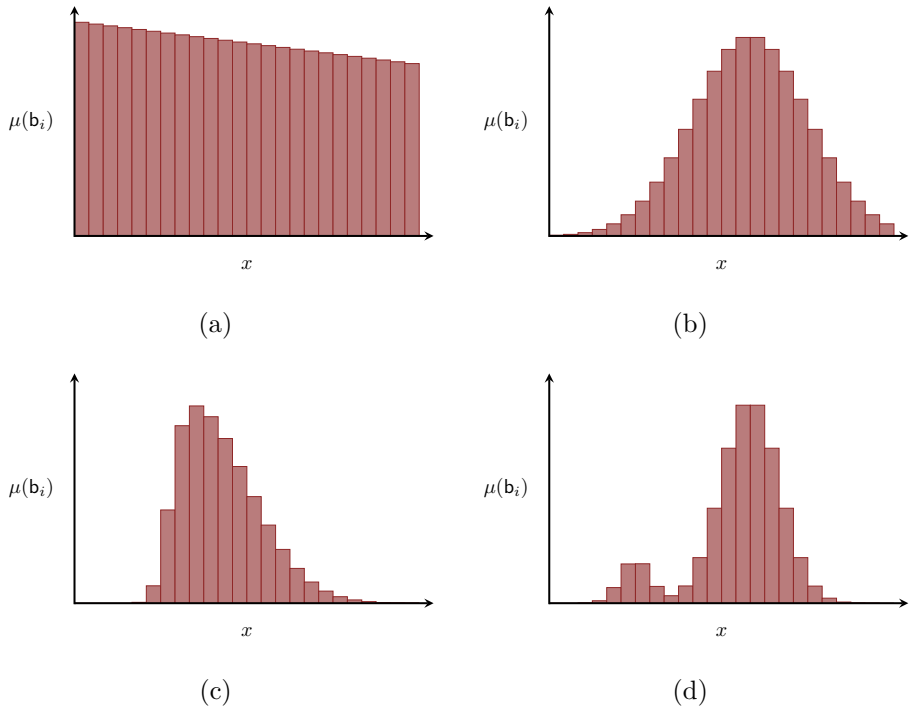

Figure 10: Histogram are extremely effective at communicating the basic features of a measure. The measure in (a) is diffuse but decaying, allocating more measure at smaller values than larger values. Conversely the measure in (b) concentrates around a single point while the measure in (c) concentrates around multiple, distinct points. Finally the measure in (d) concentrates around a single point, but that concentration is strongly asymmetric unlike the concentration in (b).

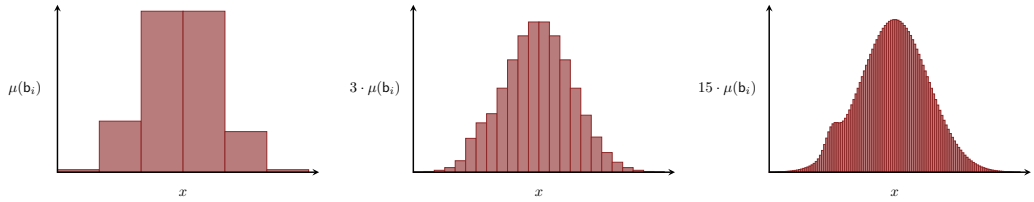

Figure 11: A histogram with a finer binning communicates more detail about a given measure, but also requires the computation of more expectation values and hence is more expensive to construct. Here as we use smaller bins we start to resolve a small side mode. Note that as we decrease the bins we also decrease the allocated measures, and hence the height of each rectangle. Here the heights are scaled to accommodate the smaller measures and make the comparison between the histograms easier.

The measure allocated to these interval subsets quantifies how the measure accumulates as we scan across the space,

$$\begin{aligned} M : X &\to [0, \mu(X)] \\ x &\mapsto M(x) = \mu(I_x). \end{aligned}$$

According this mapping is known as a **cumulative distribution function** (Figure 12).

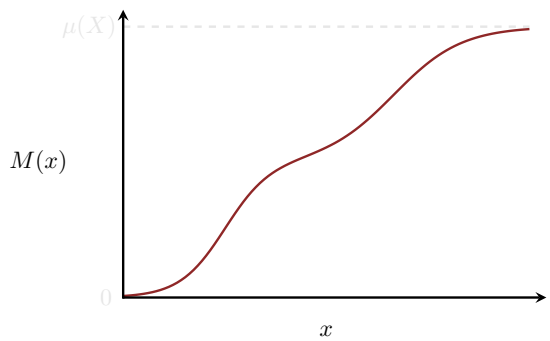

Figure 12: A cumulative distribution function quantifies how measure is allocated to expanding intervals on an ordered space. At the lower boundary of the space the interval contains no points and the cumulative distribution function returns zero. As we move towards larger values the interval expands, accumulating more and more measure. Finally at the upper boundary of the space the interval asymptotes to the total measure.

Cumulative distribution functions are also sometimes written is a way that communicates the measure and interval at the same time, for example $\mu([x' < x])$ or even $\mu(x' < x)$. Personally I find these notations to be a bit too confusing as it's easy to mistake which variable denotes points in the interval and which variable defines the upper boundary of the interval itself.

By construction if $x_1 < x_2$ then $I_{x_1} \subset I_{x_2}$. Consequently

$$\mu(I_{x_1}) \leq \mu(I_{x_2})$$

or, equivalently,

$$M(x_1) \leq M(x_2).$$

In other words every cumulative distribution function is a monotonically non-decreasing function that begins at 0 and ends at $\mu(X)$.

The precise shape of this non-decreasing accumulation conveys many features of the ambient measure. For example if the measure concentrates around a single point then the cumulative distribution function will rapidly increase around that point, increasing only slowly before and after (Figure 13a). In general the faster the cumulative distribution function increases the

stronger the concentration will be (Figure 13b). Similarly if there are any gaps in the allocation, intermediate intervals with zero allocated measure, then the cumulative distribution function will flatten out completely (Figure 13c).

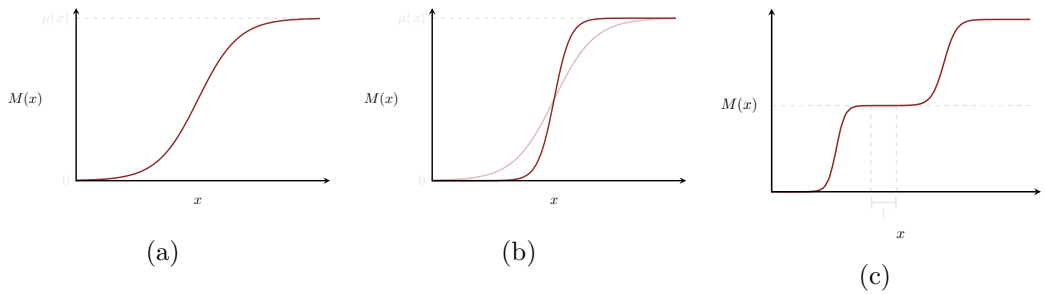

Figure 13: A careful survey of a cumulative distribution function can communicate a wealth of information about the ambient measure. (a) Here the ambient measure is unimodal with the cumulative distribution function appreciably increasingly only one we reach the central neighborhood where the measure allocation is concentrated. (b) A narrower concentration results in a steeper cumulative distribution function. (c) A cumulative distribution function flattens if there are any gaps in the measure allocation. Here the measure concentrates around two points separated by null interval I in between.

One really nice feature of cumulative distribution functions is that they allow us to compute explicit interval probabilities. The union of any two-sided, half-open interval

$$(x_1, x_2] = \{x \in X \mid x_1 < x \leq x_2\}$$

with the disjoint one-sided interval I_{x_1} defines another one-sided interval,

$$I_{x_2} = I_{x_1} \cup (x_1, x_2].$$

Measure additivity then implies that

$$\begin{aligned} \mu(I_{x_2}) &= \mu(I_{x_1} \cup (x_1, x_2]) \\ &= \mu(I_{x_1}) + \mu((x_1, x_2]) \end{aligned}$$

or

$$\begin{aligned} \mu(I_{x_2}) &= \mu(I_{x_1}) + \mu((x_1, x_2]) \\ M(x_2) &= M(x_1) + \mu((x_1, x_2]) \\ \mu((x_1, x_2]) &= M(x_2) - M(x_1). \end{aligned}$$

In words the measure allocated to any two-sided interval can be computed by subtracting the cumulative distribution function outputs at the interval boundaries (Figure 14).

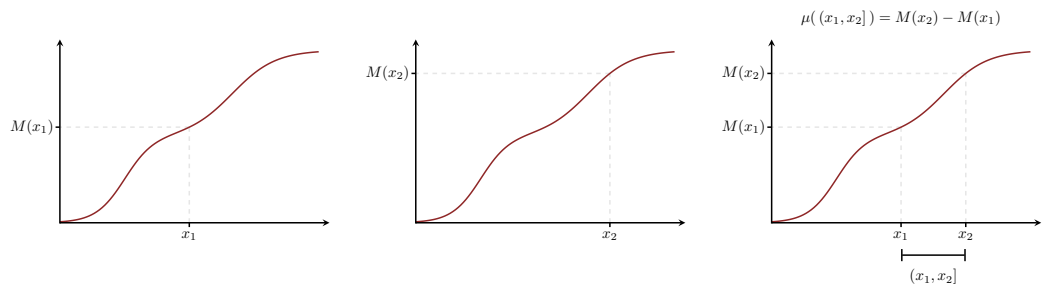

Figure 14: The difference of cumulative distribution function outputs at two points is equal to the measure allocated to the interval spanning those two points. This allows us to calculate interval measure allocations as needed.

If the measure allocated to every measurable subset $\mathsf{x} \in \mathcal{X}$ can be derived from interval allocations then a cumulative distribution function will provide enough information to compute the measure allocated to every measurable subset. In other words the cumulative distribution function in this case *completely* characterizes the measure, and can be considered as alternative way to define measures entirely. Conveniently on every ordered measurable space that we will encounter, such spaces of integers and real numbers equipped with Borel σ-algebras, this will be true.

On an ordered, countable space the cumulative distribution function can be written as the sum of mass function evaluations,

$$\begin{aligned} M(x) &= \mu(\{x' \in X \mid x' \leq x\}) \\ &= \sum_{x' \leq x} \mu(x'). \end{aligned}$$

Consequently mass functions and cumulative distribution functions provide redundant information on these spaces (Figure 15).

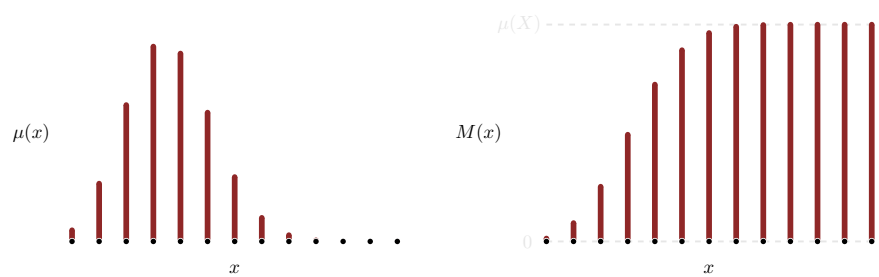

Figure 15: On ordered, countable spaces a mass function and cumulative distribution function provide equivalent, and hence redundant, characterizations of a measure.

Mass functions do not completely define a measure, however, on ordered but uncountable spaces. In this case a cumulative distribution function can provide the information that the element-wise allocations lacked.

For example on a real line a continuous cumulative distribution function defines a measure that allocates zero to every atomic subset but still manages to accumulate finite measure as we scan through the space. Any jumps in a cumulative distribution function correspond to individual elements that have been allocated non-zero measure (Figure 16).

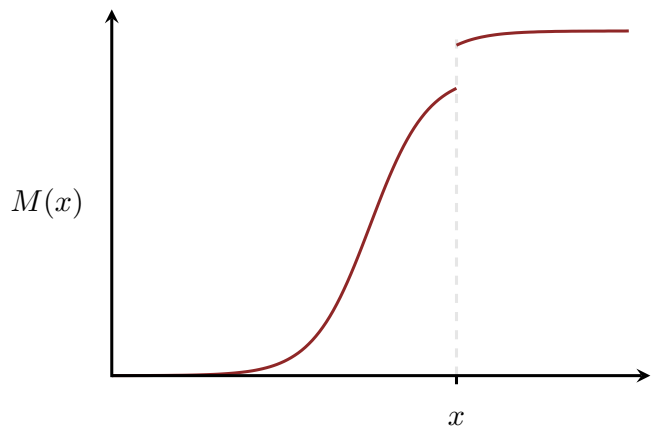

Figure 16: When the ambient space is ordered and uncountable and every atomic subset is a null subset then the cumulative distribution function will be continuous. Any discontinuities in a cumulative distribution function correspond to exceptional atomic subsets that have been allocated finite measure.

4.4 Quantiles

When a cumulative distribution function is bijective, mapping each point $x \in X$ in the ambient space to a unique accumulated measure $\mu(I_x) = M(x)$, we can invert it to map any accumulated measure to the point at which that accumulation is achieved (Figure 17),

$$\begin{aligned} q_\mu : [0, \mu(X)] &\to X \\ m \qquad &\mapsto M^{-1}(m). \end{aligned}$$

This inverse mapping is known as a **quantile function**.

Because quantiles of probability distribution functions are particularly useful in some applications they are often given explicit names. For example the point at which half of the total probability has been accumulated,

$$M(x_{0.5}) = 0.5,$$

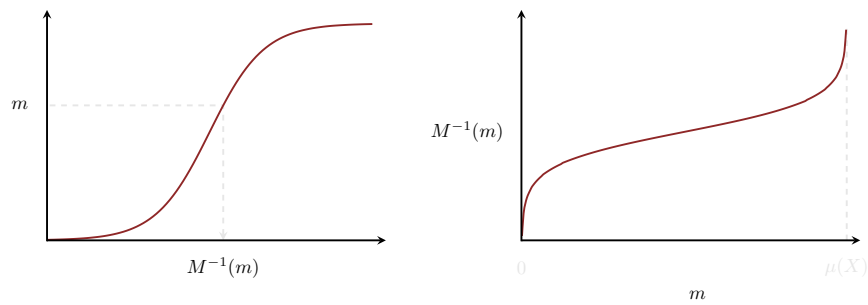

Figure 17: If a cumulative distribution function is invertible then its inverse defines a quantile function that maps accumulated measures to the points in the ambient space where the accumulation is reached.

is denoted the **median** of the probability distribution. On spaces where a mean is well-defined the median and mean complement each other by quantifying slightly different notions of centrality. Similarly the points where a quarter of the probability has been accumulated and a quarter of the probability remains,

$$M(x_{0.25}) = 0.25$$
$$M(x_{0.75}) = 0.75,$$

are known as the **quartiles**.

If the cumulative distribution function is not continuous then the quantile function will not be well-defined. For example on a countable space the cumulative distribution function can achieve only a countable number of accumulated measures. Any intermediate value m can be only bounded below by the point x_{m-} that achieves the largest accumulated measure below m,

$$x_{m-} = \underset{x \in X}{\text{argmax}} M(x) < m,$$

and bounded above by the point x_{+} that achieves the smallest accumulated measure above m (Figure 18),

$$x_{m+} = \underset{x \in X}{\text{argmin}} M(x) > m.$$

Many software packages implement quantile functions that *heuristically* interpolate between the x_{m-} and x_{m+} to provide a single value when when the cumulative distribution function is not invertible. Each interpolation strategy defines a different heuristic quantile function.

5 Algorithmic Expectation

Up to this point our discussion of general expectation values has been theoretical. Given a measure μ we have shown that a linear map from sufficiently nice, real-valued functions f to

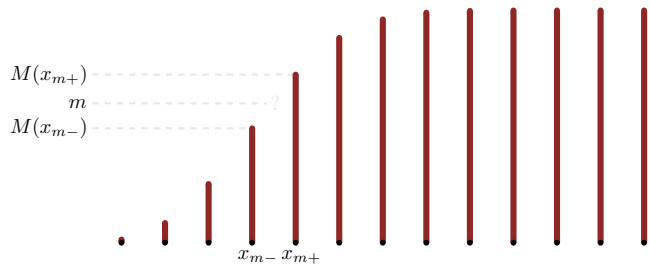

Figure 18: Cumulative distribution functions on countable spaces are not invertible. Only a countable number of measure accumulations occur at individual points; most measure accumulations occur "in between" the countable points.

real numbers $\mathbb{E}_\mu[f] \in \mathbb{R}$ exists. We do not yet know, however, how to *evaluate* that map to give explicit expectation values in practice.

Fortunately a few exceptional measures produce expectation values that can be computed from certain mathematical operations that can be implemented algorithmically, if not exactly then approximately. In this section we'll review these exceptional measures and their practical consequences.

5.1 Expectation on Countable Spaces

Because they can be completely specified by a mass function, measure allocations on countable spaces are particularly straightforward to implement in practice. Conveniently expectation values on these spaces are also completely specified by mass functions.

5.1.1 Expectation As Summation

For any countable measurable space $(X, 2^X)$ we can always decompose a real-valued function into a sum of atomic indicator functions,

$$f(x) = \sum_{x' \in X} f(x') \cdot I_{\{x'\}}(x).$$

The expectation value with respect to any measure μ follows immediately by applying linearity,

$$\begin{aligned} \mathbb{E}_\mu[f] &\equiv \mathbb{E}_\mu \left[\sum_{x' \in X} f(x') \cdot I_{\{x'\}} \right] \\ &= \sum_{x' \in X} f(x') \cdot \mathbb{E}_\mu \left[I_{\{x'\}} \right], \end{aligned}$$

and then the definition of expectation values for indicator functions,

$$\begin{aligned}\mathbb{E}_\mu[f] &= \sum_{x' \in X} f(x') \cdot \mathbb{E}_\mu \left[I_{\{x'\}} \right] \\ &= \sum_{x' \in X} f(x') \cdot \mu(\{x'\}).\end{aligned}$$

Consequently expectation values with respect to *any* real-valued function on these measure spaces reduces to explicit summations. Moreover the summations are informed by only the measure allocations to atomic subsets, which is exactly what a mass function specifies. In particular if X is finite the general definition of expectation value reduces to the heuristic construction that we considered in Section 1.

5.1.2 Practical Consequences

When X if finite we can implement these expectation value summations by directly looping over each expectand outputs. Unfortunately this approach becomes unfeasible if X contains a countably infinite number of elements.

Some infinite sums do enjoy closed-form solutions. For all other sums we cannot evaluate the corresponding expectation values exactly. That said we may be able to *approximate* them: summing over only a finite number of elements with both $\mu(x)$ and $f(x)$ large enough can give answers close to the exact expectation values.

Consider, for example, the counting measure that allocates unit measure to each atomic subset,

$$\chi(\{x\}) = 1.$$

More generally the counting measure allocates measure by counting the number of elements contained in a given subset,

$$\chi \mathsf{x} = \sum_{x \in \mathsf{x}} 1.$$

The expectation value of any real-valued function $f : X \to \mathbb{R}$ with respect to counting measure is given by over summing all of the output values,

$$\begin{aligned}\mathbb{E}_\chi[f] &= \int \chi(\mathrm{d}x)\, f(x) \\ &= \sum_{x \in X} \chi(x) \cdot f(x) \\ &= \sum_{x \in X} 1 \cdot f(x) \\ &= \sum_{x \in X} f(x).\end{aligned}$$

In other words all expectation values with respect to the counting measure can be implemented by simply summing over the expectand outputs.

We can scale the counting measure by a positive function $g : X \to \mathbb{R}^+$ following the strategy introduced in Section 3.2. The scaled measure $g \cdot \chi$ is implicitly defined by the expectation values

$$\begin{aligned}
\mathbb{E}_{g \cdot \chi}[f] &= \mathbb{E}_{\chi}[g \cdot f] \\
&= \sum_{x \in X} \chi(x) \cdot (g(x) \cdot f(x)) \\
&= \sum_{x \in X} (g(x) \cdot \chi(x)) \cdot f(x).
\end{aligned}$$

Consequently $g \cdot \chi$ can be implemented by simply scaling the element-wise allocations,

$$(g \cdot \chi)(x) = g(x) \cdot \chi(x).$$

While this might have seemed obvious from the beginning, the machinery of expectation values allows us to prove that this intuitive definition is consistent with how measure theory behaves more generally.

5.2 Lebesgue Expectation on Real Lines

Frustratingly there are no universal strategies for directly evaluating expectation values on uncountable spaces. Sometimes, however, the structure of an uncountable ambient space allow us to reduce expectation values to more feasible mathematical operations. In particular expectation values with respect to the Lebesgue measure on a real line can be related to the familiar integral from calculus.

5.2.1 Expectation As Integration

By definition the expectation value of an indicator function is given by the measure allocated to the defining subset,

$$\mathbb{E}_{\mu}[I_{\mathsf{x}}] = \mu(\mathsf{x}).$$

On a real line $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ the Lebesgue measure allocated to any interval is just the distance between the end points,

$$\lambda(\,[x_1, x_2]\,) = d(x_1, x_2) = |x_2 - x_1|.$$

Consequently the Lebesgue expectation value of an interval indicator function becomes

$$\mathbb{E}_{\lambda}[I_{[x_1, x_2]}] = |x_2 - x_1|.$$

That expectation value, however, also happens to be the area under the curve defined by the corresponding indicator function (Figure 19a),

$$\begin{aligned}\text{area} &= \text{height} \cdot \text{length} \\ &= 1 \cdot |x_2 - x_1| \\ &= \mathbb{E}_\lambda[I_{[x_1,x_2]}].\end{aligned}$$

This geometric coincidence also generalizes to simple functions. The area under the curve defined by a simple function built from a single interval

$$s(x) = \phi \cdot I_{[x_1,x_2]}$$

is just

$$\begin{aligned}\text{area} &= \text{height} \cdot \text{length} \\ &= \phi \cdot |x_2 - x_1| \\ &= \phi \cdot \mathbb{E}_\lambda[I_{[x_1,x_2]}] \\ &= \mathbb{E}_\lambda[\phi \cdot I_{[x_1,x_2]}] \\ &= \mathbb{E}_\lambda[s].\end{aligned}$$

More generally the area under the curve defined by a simple function built from many intervals

$$s(x) = \sum_j \phi_j \cdot I_{[x_{1,j},x_{2,j}]}$$

is built up from rectangles defined by each component,

$$\begin{aligned}\text{area} &= \sum_j \text{area}_j \\ &= \sum_j \text{height}_j \cdot \text{length}_j \\ &= \sum_j \phi_j \cdot |x_{2,j} - x_{1,j}| \\ &= \sum_j \mathbb{E}_\lambda[\phi_j \cdot I_{[x_{1,j},x_{2,j}]}].\end{aligned}$$

By linearity, however, this is just the expectation value of the simple function itself (Figure 19b)

$$\begin{aligned}\text{area} &= \sum_j \mathbb{E}_\lambda[\phi_j \cdot I_{[x_{1,j},x_{2,j}]}] \\ &= \mathbb{E}_\lambda[\sum_j \phi_j \cdot I_{[x_{1,j},x_{2,j}]}] \\ &= \mathbb{E}_\lambda[s].\end{aligned}$$

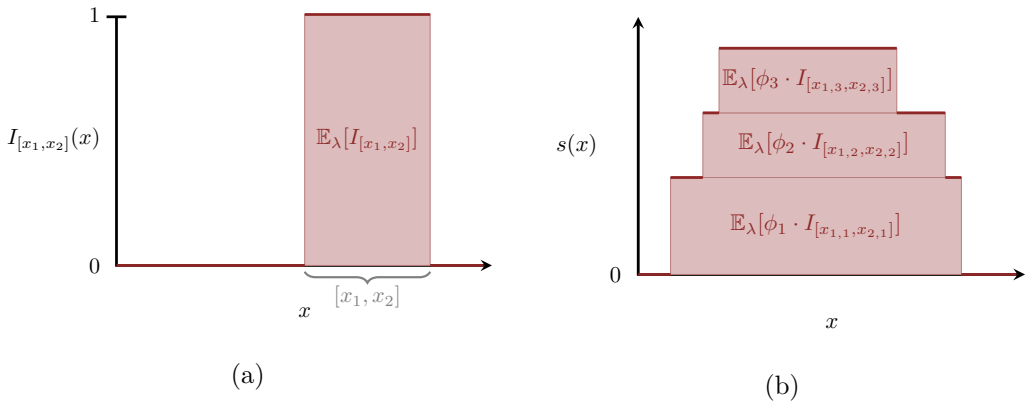

Figure 19: Expectations of simple functions with respect to the Lebesgue measure are intimately related to the area under the curve defined by simple functions. (a) The area under the curve defined by an interval indicator function is equal to the height, 1, times the length of the interval. That, however, is just equal to the Lebesgue expectation of the indicator function itself. (b) The area under the curve defined by interval simple functions is built up from the area of rectangles from each component. The total area is equal to the expectation value of the simple function itself.

Decomposing the positive and negative parts of a measurable, real-valued function into simple functions pushes this relationship further. On one hand we can use the decomposition to define general expectation values, and on the other we can use it to compute the area under the curve defined by any sufficiently nice function (Figure 20a).

We can also use calculus to compute the same area under the curve. A Riemann integral is defined by partitioning the real line into equally-sized intervals and then constructing rectangles from the height of the integrand at the end of each interval. As the intervals become smaller and smaller the sum of the rectangle areas converges to an integral (Figure 20b),

$$\int \mathrm{d}x\, f(x) = \lim_{\delta \to 0} \sum_{n=-\infty}^{\infty} delta \cdot f(x_0 + n \cdot \delta).$$

Geometrically Lebesgue expectation values compute the area under a curve by summing over *vertically* stacked rectangles while Riemann integration computes the area by summing over *horizontally* stacked rectangles. Riemann integration doesn't always result in a well-defined answer, but when it does we can use these two methods for computing the area under a curve to relate expectation with respect to the Lebesgue measure to classic integration! More formally

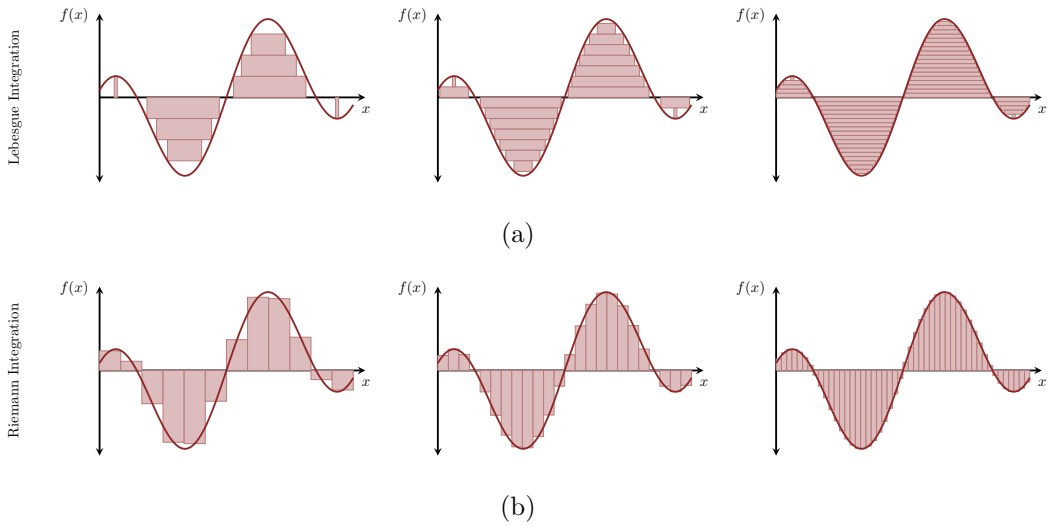

Figure 20: On a real line $X = \mathbb{R}$ expectation with respect to the Lebesgue measure and Riemann integration both quantify the area under a curve defined by a sufficiently nice real-valued function $f : \mathbb{R} \to \mathbb{R}$. (a) As we add more components the expectation value of a simple function converges to the Lebesgue expectation value of f. At the same time the sum of the rectangular areas defined by each component converges to the area under the curve defined by f. (b) Riemann integration also computes the area under the curve as a sum of increasingly narrow rectanglular areas, only the rectangles are stacked horizontally instead of vertically.

for any measurable, real-valued function $f\mathbb{R} \to \mathbb{R}$ we have

$$\begin{aligned}
\mathbb{E}_\lambda[f] &= \int \lambda(\mathrm{d}x)\, f(x) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x\, f(x)
\end{aligned}$$

so long as the integral $\int \mathrm{d}x\, f(x)$ is well-defined.

This particular equivalence is one motivation for the many alternative expectation value notations that we discussed in Section 2.4. In general the integrals in those notations do not correspond to calculus integrals but in the special case of the Lebesgue measure over a real line they do!

5.2.2 Practical Consequences

When a real-valued function has a well-defined integral then we can apply the tools of calculus to evaluate Lebesgue expectation values. The exceptional integrals that can be evaluated analytically allow us to compute the corresponding Lebesgue expectation values exactly. More generally we often have to resort to numerical integration techniques to approximate integrals, and hence approximately evaluate Lebesgue expectation values.

For example the expectation value of an interval indicator function is given by

$$\begin{aligned}
\mathbb{E}_\lambda[I_{[x_1,x_2]}] &= \int_{-\infty}^{\infty} \mathrm{d}x\, I_{[x_1,x_2]}(x) \\
&= \int_{x_1}^{x_2} \mathrm{d}x \\
&= x_2 - x_1,
\end{aligned}$$

consistent with the definition of the Lebesgue measure. Note that the correct, positive answer required that we integrate from the lower end of the interval to the upper end. Changing the order defines the same interval, and hence the same Lebesgue measure allocation, but it flips the sign of the calculus integral. In order to properly relate Lebesgue expectation values to integrals we have to fix the *orientation* of the integrals.

Similarly the mean would by given by the integral of the identity function,

$$\begin{aligned}
\mathbb{E}_\lambda[\iota] &= \int_{-\infty}^{\infty} \mathrm{d}x\, \iota(x) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x\, x \\
&= \infty - \infty.
\end{aligned}$$

Unfortunately this results in an ill-posed answer because ∞ minus itself is consistent with *every* value on the real line. Had we been a bit more careful, however, this would not have been surprising. The problem is that the identify function is not Lebesgue-integrable,

$$\begin{aligned}
\mathbb{E}_{\lambda}[|\iota|] &= \int_{-\infty}^{\infty} \mathrm{d}x \, |\iota(x)| \\
&= 2 \int_{0}^{\infty} \mathrm{d}x \, x \\
&= \infty!
\end{aligned}$$

Consequently the Lebesgue measure does not have any well-defined moments.

Likewise the scaling of the Lebesgue measure by a positive function $g : X \to \mathbb{R}^{+}$ can be implemented with the integrals

$$\begin{aligned}
\mathbb{E}_{g \cdot \lambda}[f] &= \mathbb{E}_{\lambda}[g \cdot f] \\
&= \int_{-\infty}^{+\infty} \mathrm{d}x \, g(x) \cdot f(x).
\end{aligned}$$

In particular the measure allocated to any interval becomes

$$\begin{aligned}
(g \cdot \lambda)(\, [x_1, x_2] \,) &= \mathbb{E}_{g \cdot \lambda}[I_{[x_1, x_2]}] \\
&= \mathbb{E}_{\lambda}[g \cdot I_{[x_1, x_2]}] \\
&= \int_{-\infty}^{+\infty} \mathrm{d}x \, g(x) \cdot I_{[x_1, x_2]}(x) \\
&= \int_{x_1}^{x_2} \mathrm{d}x \, g(x).
\end{aligned}$$

6 Conclusion

Expectation values are the main way that we interact with measures and probability distributions, both in theory and in practice. Indeed a recurring theme in applying probability theory in practice will be the principled computation of expectation values for relevant expectands.

In the next chapter we'll learn how to apply the exceptional algorithmic expectation values with respect to the Lebesgue measures to a much larger class of measures. Later on we'll learn some powerful *sampling* techniques for estimating general expectation values directly.

Acknowledgements

A very special thanks to everyone supporting me on Patreon: Adam Fleischhacker, Adriano Yoshino, Alan Chang, Alessandro Varacca, Alexander Bartik, Alexander Noll, Alexander Petrov, Alexander Rosteck, Anders Valind, Andrea Serafino, Andrew Mascioli, Andrew Rouillard, Andrew Vigotsky, Angie_Hyunji Moon, Ara Winter, Austin Rochford, Austin Rochford, Avraham Adler, Ben Matthews, Ben Swallow, Benjamin Glemain, Bradley Kolb, Brandon Liu, Brynjolfur Gauti Jónsson, Cameron Smith, Canaan Breiss, Cat Shark, Charles Naylor, Chase Dwelle, Chris Jones, Chris Zawora, Christopher Mehrvarzi, Colin Carroll, Colin McAuliffe, Damien Mannion, Damon Bayer, dan mackinlay, Dan Muck, Dan W Joyce, Dan Waxman, Dan Weitzenfeld, Daniel Edward Marthaler, Darshan Pandit, Darthmaluus , David Burdelski, David Galley, David Wurtz, Denis Vlašiček, Doug Rivers, Dr. Jobo, Dr. Omri Har Shemesh, Ed Cashin, Edgar Merkle, Eric LaMotte, Erik Banek, Ero Carrera, Eugene O'Friel, Felipe González, Fergus Chadwick, Finn Lindgren, Florian Wellmann, Francesco Corona, Geoff Rollins, Greg Sutcliffe, Guido Biele, Hamed Bastan-Hagh, Haonan Zhu, Hector Munoz, Henri Wallen, hs, Hugo Botha, Håkan Johansson, Ian Costley, Ian Koller, idontgetoutmuch, Ignacio Vera, Ilaria Prosdocimi, Isaac Vock, J, J Michael Burgess, Jair Andrade, James Hodgson, James McInerney, James Wade, Janek Berger, Jason Martin, Jason Pekos, Jason Wong, Jeff Burnett, Jeff Dotson, Jeff Helzner, Jeffrey Erlich, Jesse Wolfhagen, Jessica Graves, Joe Wagner, John Flournoy, Jonathan H. Morgan, Jonathon Vallejo, Joran Jongerling, Joseph Despres, Josh Weinstock, Joshua Duncan, JU, Justin Bois, Karim Naguib, Karim Osman, Kejia Shi, Kristian Gårdhus Wichmann, Kádár András, Lars Barquist, lizzie , LOU ODETTE, Marc Dotson, Marcel Lüthi, Marek Kwiatkowski, Mark Donoghoe, Markus P., Martin Modrák, Matt Moores, Matt Rosinski, Matthew, Matthew Kay, Matthieu LEROY, Maurits van der Meer, Merlin Noel Heidemanns, Michael Colaresi, Michael DeWitt, Michael Dillon, Michael Lerner, Mick Cooney, Márton Vaitkus, N Sanders, Name, Nathaniel Burbank, Nic Fishman, Nicholas Clark, Nicholas Cowie, Nick S, Nicolas Frisby, Octavio Medina, Oliver Crook, Olivier Ma, Patrick Kelley, Patrick Boehnke, Pau Pereira Batlle, Peter Smits, Pieter van den Berg , ptr, Putra Manggala, Ramiro Barrantes Reynolds, Ravin Kumar, Raúl Peralta Lozada, Riccardo Fusaroli, Richard Nerland, Robert Frost, Robert Goldman, Robert kohn, Robin Taylor, Ross McCullough, Ryan Grossman, Rémi , S Hong, Scott Block, Sean Pinkney, Sean Wilson, Seth Axen, shira, Simon Duane, Simon Lilburn, sssz, Stan_user, Stefan, Stephanie Fitzgerald, Stephen Lienhard, Steve Bertolani, Stew Watts, Stone Chen, Susan Holmes, Svilup, Sören Berg, Tao Ye, Tate Tunstall, Tatsuo Okubo, Teresa Ortiz, Thomas Lees, Thomas Vladeck, Tiago Cabaço, Tim Radtke, Tobychev , Tom McEwen, Tony Wuersch, Utku Turk, Virginia Fisher, Vitaly Druker, Vladimir Markov, Wil Yegelwel, Will Farr, Will Tudor-Evans, woejozney, yolhaj , Zach A, Zad Rafi, and Zhengchen Cai.

License