Measured-Informed Integration and Expectation

Michael Betancourt

August 2023

Table of contents

6 Conclusion   41

Acknowledgements   41

License   42

In Chapter Four 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 numbers that generalizes Riemann integration from calculus. This measure-informed integration 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 measure-informed integration 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 measure-informed integrals can be used to implicitly define measures without having to explicitly define subset allocations and some useful applications of these implicit specifications. Finally we'll consider particular measure-informed integrals that are distinguished by common ambient space structures and then conclude with a discussion of a few exceptional measures whose integrals can be computed algorithmically.

1 Integration on Finite Measure Spaces

To start our discussion of measure-informed integration as simply as possible let's begin by considering a finite measure space compromised of the finite set

$$X = \{\square, \clubsuit, \diamondsuit, \heartsuit, \spadesuit\},$$

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 weight 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 $f(x)$,

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

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

$$\sum_{x \in X} \mu(x) \cdot f(x) = \quad \mu(\square) \cdot f(\square)$$
$$+ \mu(\clubsuit) \cdot f(\clubsuit)$$
$$+ \mu(\diamondsuit) \cdot f(\diamondsuit)$$
$$+ \mu(\heartsuit) \cdot f(\heartsuit)$$
$$+ \mu(\spadesuit) \cdot f(\spadesuit).$$

This summary emphasizes not only large output values but also outputs from highly-weighted inputs. For example even if the output $f(\square)$ is small the contribution from $\square$ can still be important if the atomic allocation $\mu(\square)$ is large.

This summary defines the **integral** of f with respect to μ,

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

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

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

$$\mathbb{I}_\mu[\alpha \cdot f + \beta \cdot g] = \sum_{x \in X} \mu(x) \cdot [\alpha \cdot f(x) + \beta \cdot g(x)]$$
$$= \sum_{x \in X} \mu(x) \cdot (\alpha \cdot f(x) + \beta \cdot g(x))$$
$$= \alpha \cdot \sum_{x \in X} \mu(x) \cdot f(x) + \beta \cdot \sum_{x \in X} \mu(x) \cdot g(x)$$
$$= \alpha \cdot \mathbb{I}_\mu[f] + \beta \cdot \mathbb{I}_\mu[g].$$

We will exploit this linearity property *endlessly* when working with measure-informed integrals.

The measure-informed integral $\mathbb{I}_\mu[f]$ 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. Measure-informed integrals 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 measure-informed integral $\mathbb{I}_{\mu_j}[f]$ as encoding the answer (Figure 1a).
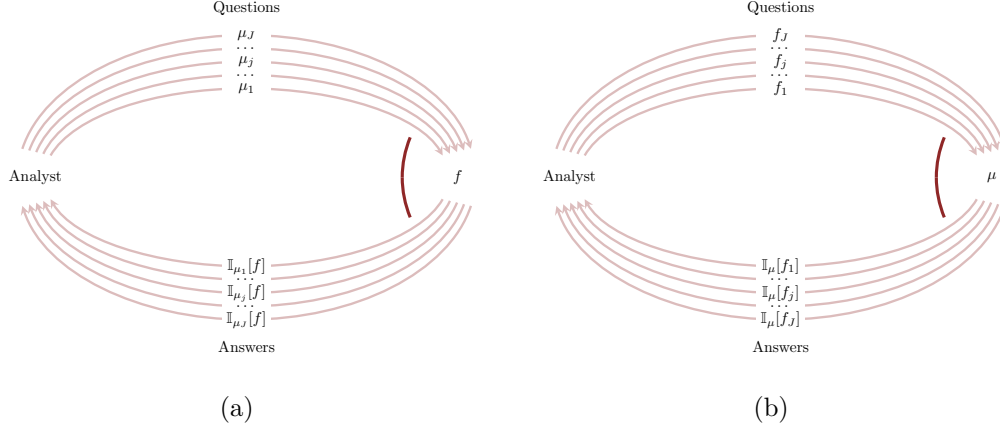
3

Figure 1: Measure-informed integrals probe the interaction between a measure and a real-valued function. (a) If we fix the function f then measure-informed integrals 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 measure-informed integral providing an answer. (b) Similarly measure-informed integrals 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 measure-informed integral encoding an answer.

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

$$\delta_{x'}(x) = \left\{ \begin{array}{ll} 1, & x = x' \\ 0, & x \neq x' \end{array} \right. .$$

The measure-informed integrals of any real-valued function f with respect to δ_{x_i} are given by

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

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

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

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

4

The corresponding measure-informed integrals captures the **average** of the function output values,

$$\mathbb{I}_\pi[f] = \sum_{x \in X} \pi(x) \cdot f(x)$$

$$= \sum_{x \in X} \frac{1}{I} \cdot f(x)$$

$$= \frac{1}{I} \sum_{x \in X} f(x).$$

When we use non-uniform measures this measure-informed integration 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. Measure-informed integrals 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 measure-informed integrals $\mathbb{I}_\mu[f_j]$ 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,

$$I_\mathsf{x}(x) = \left\{ \begin{array}{ll} 1, & x \in \mathsf{x} \\ 0, & x \notin \mathsf{x} \end{array} \right. .$$

In other words an indicator function *indicates* whether or not a point is contained in the defining subset.

The measure-informed integral of the indicator function I_x, however, is just the measure allocated to x,

$$\mathbb{I}_\mu[I_\mathsf{x}] = \sum_{x \in X} \mu(x) \cdot I_\mathsf{x}(x)$$

$$= \sum_{x \in \mathsf{x}} \mu(x) \cdot I_\mathsf{x}(x) + \sum_{x \notin \mathsf{x}} \mu(x) \cdot I_\mathsf{x}(x)$$

$$= \sum_{x \in \mathsf{x}} \mu(x) \cdot 1 + \sum_{x \notin \mathsf{x}} \mu(x) \cdot 0$$

$$= \sum_{x \in \mathsf{x}} \mu(x)$$

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

Measure-informed integrals of various indicator functions allow us to directly probe the various subset allocations.

2 Integration on General Measure Spaces

Unfortunately the straightforward construction of measure-informed integrals 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 measure-informed integrals to arbitrary measure spaces we have to appeal to a more sophisticated construction with some subtle, but important, consequences.

2.1 Integration of Simple Functions

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

For example consider indicator functions which vanish outside of a given measurable subset (Figure 2)

$$I_{\mathsf{x}}(x) = \left\{ \begin{array}{ll} 1, & x \in \mathsf{x} \\ 0, & x \notin \mathsf{x} \end{array} \right. .$$
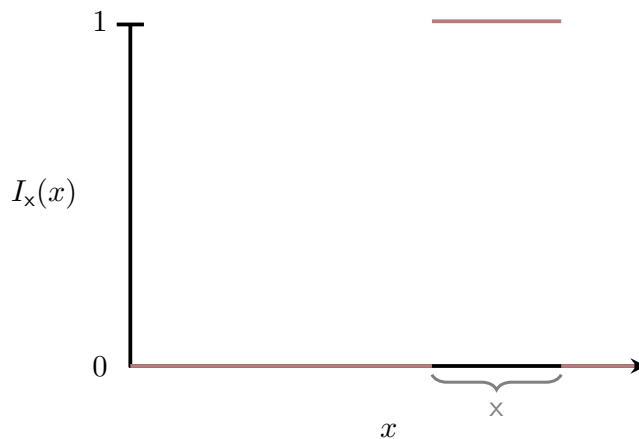

Figure 2: An indicator function corresponding to a measurable subset $\mathsf{x} \in \mathcal{X}$ vanishes for all inputs that are not contained in x. Here x is an interval subset over a real line.

In order to generalize the behavior on finite measure spaces that we encountered in the measure-informed integral of any indicator function should be equal to the measure

allocated to that subset,

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

for any measure μ.

We can manufacture even more complex functional behavior 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.

If we assume that measure-informed integration is a linear operation on any measure space then we can immediately compute the integral of any simple function,

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

2.2 Integration of Measurable Functions

Most functions whose measure-informed integrals are of interest in practical analysis are not simple functions. Relevant functions, however, can often be *well-approximated* by simple functions. As we incorporate more and more indicator functions we can construct simple functions that approximate non-simple functions better and better (Figure 4).

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 $(X, \mathcal{X})$,

$$f^*((-\infty, x]) = \{x \in X \mid f(x) \in (\infty, x]\} \in \mathcal{X}.$$
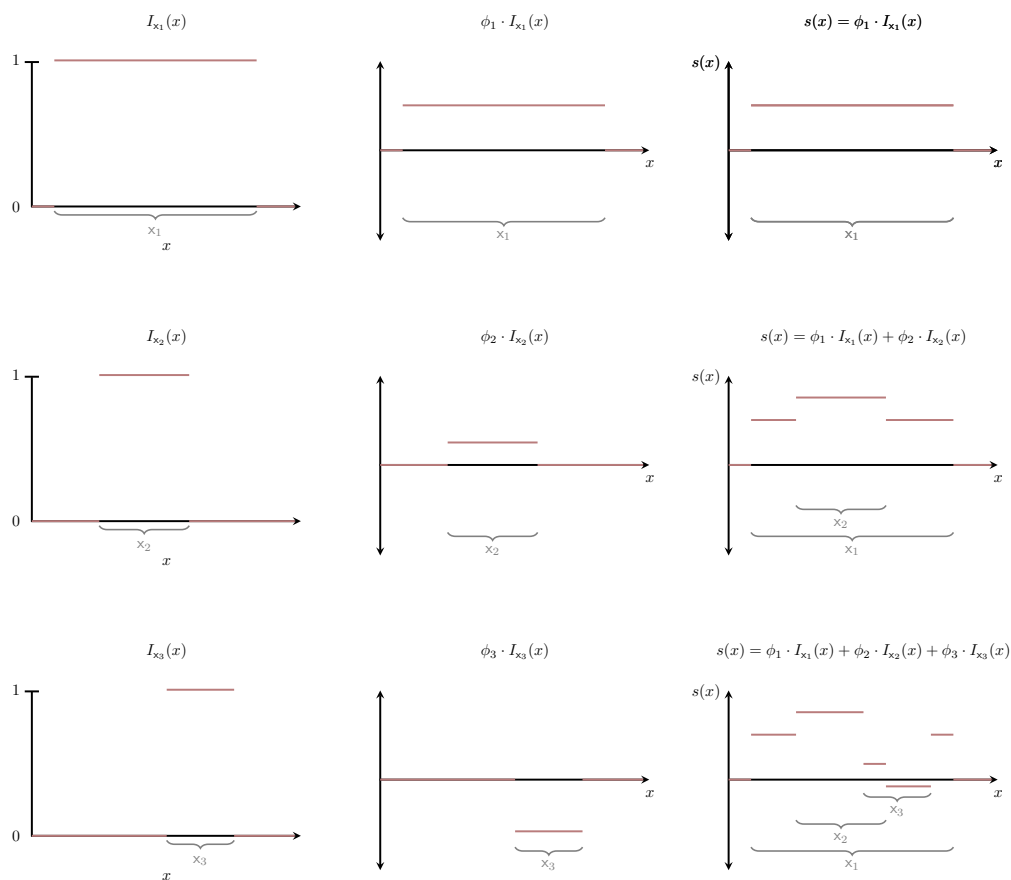
7

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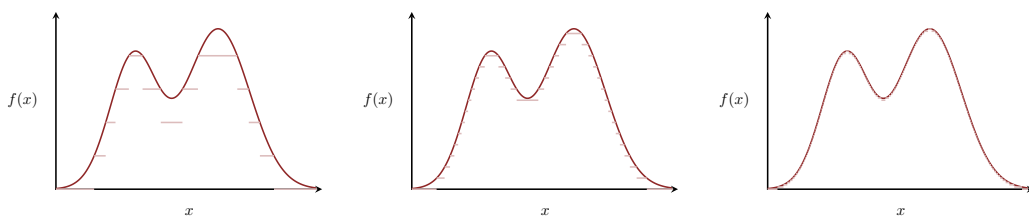

Figure 4: As we incorporate more indicator functions simple functions become more flexible and are able to better approximate the behavior of non-simple functions. Certain non-negative functions can be *exactly* recovered from sufficiently flexible simple functions.

In practice non-measurable functions are as difficult to construct as non-measurable subsets and measurability can largely be taken for granted.

We'll come back to the topic of measurable functions in much more detail in Chapter Seven. For now our main concern will be to avoid 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 measure-informed integral of a non-negative, measurable function as the measure-informed integral of the corresponding simple function decomposition,

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

In general a non-negative, measurable function can be represented by more than one simple function decomposition. Fortunately the measure-informed integral derived from any of them will always be the same. Consequently there's no worry for ambiguous of otherwise inconsistent answers, and measure-informed integrals for non-negative, measurable function are completely well-behaved.

This procedure for defining measure-informed integrals through simple functions representations is known as **Lebesgue integration** in the mathematics literature. In this book I will use the more explicit "measure-informed integration" instead.

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 measure-informed integrals for measurable functions that aren't necessarily positive we just have to decompose the functions by the sign of their outputs (Figure 5),

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

9

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

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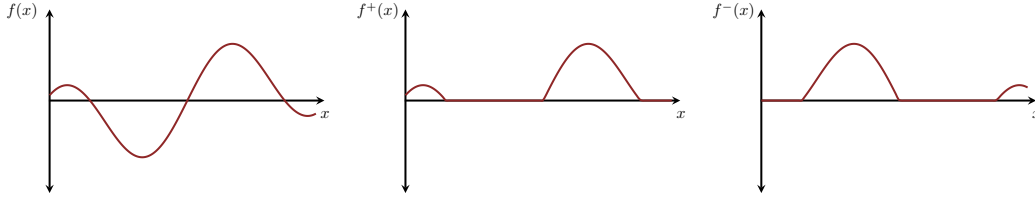$$f^-(x) = \begin{cases} -f(x), & f(x) < 0 \\ 0, & f(x) \geq 0 \end{cases}.$$

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 $f^- : X \to \mathbb{R}^+$.

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

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

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

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

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

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

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

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 **integrand**.

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. That said there are enough exceptions that we have to be careful to explicitly validate integrability in practice.

2.3 Equivalent Integrands

One subtle but important consequence of this general definition of measure-informed integration is that many integrands will yield the same measure-informed integrals 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 measure-informed integrals of these two simple functions are then related to each other by

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

When $\phi' \neq 0$ then $s(x)$ and $s'(x)$ will differ for all $x \in \mathsf{x}'$; 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 measure-informed integrals 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-integrals.

More generally any two integrands $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ will share the same μ-integrals 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

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

Intuitively modifying integrands on sets of measure zero does not affect their measure-informed integrals.

If the subset of deviant inputs is contained within a μ-null subset then 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

11

we can say that the two integrands are equal **up to subsets of measure zero** or equal **up to null subsets**.

Much of the mathematics literature overloads the equals sign when referring to measurable functions that are equal almost everywhere in equations, as that is the only notion of equals that is relevant in measure and probability theory. In this book I will be more explicit and use

$$f \stackrel{\mu}{=} g$$

whenever comparing two measurable functions that are equal up to μ-null subsets.

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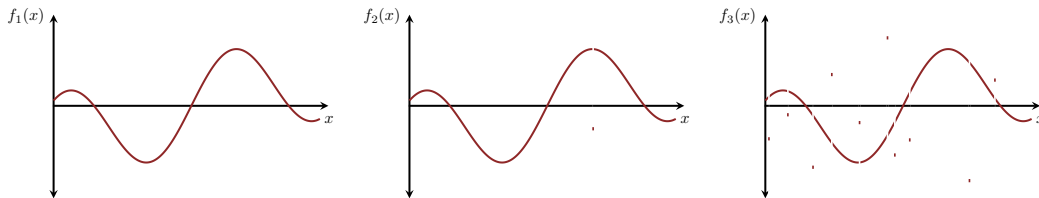

Figure 6: Because any countable collection of points is allocated zero Lebesgue measure any integrands whose outputs differ only at a countable number of inputs will yield the same measure-informed integrals. Here $\mathbb{I}_\lambda[f_1] = \mathbb{I}_\lambda[f_2] = \mathbb{I}_\lambda[f_3]$ and, from perspective of the Lebesgue measure, these functions are equivalent. In this case we write $f_1 \stackrel{\lambda}{=} f_2 \stackrel{\lambda}{=} f_3$.

Applications of measure theory can't distinguish between integrands that are equal up to sets of measure zero. If we want to avoid this ambiguity then we have to impose structural constraints to isolate a single, unique integrands from the collection of equivalent integrands. For example we can modify a continuous integrands on input subsets of measure zero without changing the measure-informed integrals, but those modifications will also introduce discontinuities. Amongst all of the equivalent integrands only one will be continuous and even though general integrands are not unique continuous integrands 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 Measure-Informed Integral Notations

One of the limitations of the measure-informed integral notation, $\mathbb{I}_\mu[f]$, 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

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

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 measure-informed integrals as

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

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 Riemann integral from calculus. We'll discuss the subtle relationship between measure-informed integration and Riemann integration in 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 measure-informed integrals as

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

or

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

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

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

and

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

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

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

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

2.5 Expectation Values

In this book we will ultimately be interested in not general measures but rather probability distributions. General measures will be used only as tools to help implement probability distributions in practice.

Within this context measure-informed integration is also known as **expectation**, with measure-informed integrals known as **expectation values**, $\mathbb{E}_\pi[f]$. Similarly integrands become **expectands**.

Technically either terminology is correct when referring to probability distributions, but I will use the expectation terminology as much as possible to better relate to the statistics literature where it is typical.

3 Specifying Measures With Integrals

To this point we have derived measure-informed integrals as a consequence of measurable subset allocations. Measure-informed integrals, however, can also be used to define measures directly, with subset allocations derived indirectly. While a bit more abstract than our initial approach this perspective does have its benefits.

3.1 Functional Perspective of Measures

Measure-informed integrals 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

$$\mathbb{I}_\mu : C(X) \to \mathbb{R}$$
$$f \mapsto \mathbb{I}_\mu[f].$$

Unfortunately this isn't technically correct because not every real-valued function has a well-defined measure-informed integral. 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 integrands.

With this notation measure-informed integration can be interpreted as a map from integrands to real numbers,

$$\mathbb{I}_{\mu} : L(X, \mathcal{X}, \mu) \to \mathbb{R}$$
$$f \qquad \mapsto \mathbb{I}_{\mu}[f].$$

In fact measure-informed integration is the only *linear* mapping 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 integrands that are not indicator functions, and hence quite a bit of redundant information about μ.

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

The integration 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 integral 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 measure-informed integrals of simple functions with respect to $\alpha \cdot \mu$ can be recovered as measure-informed integrals of *scaled* integrands with respect to μ,

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

Because general measure-informed integrals are derived from the measure-informed integrals of simple functions we will then have

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

for *every* integrand $f : X \to \mathbb{R}$. In other words these modified integrals 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 integrals

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

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

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

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

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

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

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

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

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

$$(g \cdot \pi)(X) = \mathbb{I}_{g \cdot \pi}[I_X]$$
$$= \mathbb{I}_{g \cdot \pi}[1]$$
$$= \mathbb{E}_{\pi}[g]$$

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

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

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

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

In this case the modified subset allocations become

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

Specifically we will always have

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

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,

$$(I_{\mathsf{x}'} \cdot \pi)(X) = \mathbb{0}_{I_{\mathsf{x}'} \cdot \pi}[I_X]$$
$$= \mathbb{0}_{I_{\mathsf{x}'} \cdot \pi}[1]$$
$$= \mathbb{0}_\pi[I_{\mathsf{x}'}]$$
$$= \pi(\mathsf{x}').$$

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

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

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 Integrals

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

4.1 Moments and Cumulants

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

4.1.1 Embeddings

In order for an ambient space X to be compatible with a real line it needs to share the metric structure of the real line. We say that we can **embed** a metric space X into a real line if we can construct an isometric injection $\iota : X \to \mathbb{R}$, in other words 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 by definition.

Note that the construction of an embedding requires that we fix the output structure of the output real line, specifying either a particular rigid real line or a particular parameterization of a flexible real line. An ambient might embed into *a* real line, but it cannot embed into *all* real lines at the same time.

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 a 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).
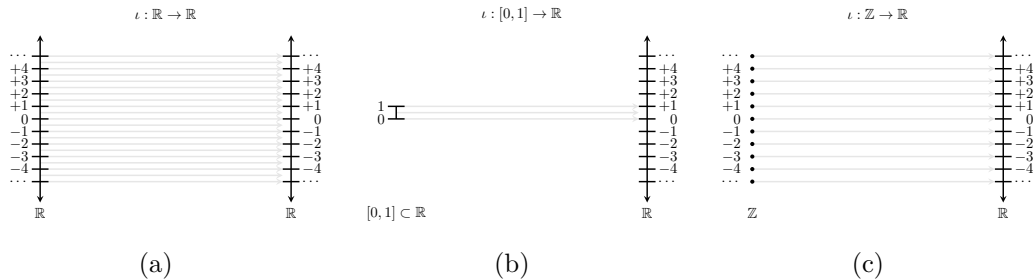

Figure 7: Many spaces naturally embed into a real line, including (a) that real line, (b) intervals of that real line, and (c) integers.

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 that subset. 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 integrand. Most embedding maps are measurable but integrability is less dependable, and failures of integrability are important in practice.

4.1.2 The Mean

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

Interpreting $\mathbb{I}_\mu[\iota]$ 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 $\mathbb{I}_\mu[\iota]$ as a weighted sum of positions,

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

The more measure that is allocated to an element the more strongly the measure-informed integral is pulled towards the position of that element. In other words $\mathbb{I}_\mu[\iota]$ 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 measure-informed integrals as simple weighted sums. Instead consider a baseline position $r_0 \in \mathbb{R}$ and the squared distance function

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

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 integrand. The measure-informed integral

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

quantifies how diffusely the measure μ is allocated around r_0; the larger the integral the less μ concentrates around r_0.

Consequently the baseline position $r_0 \in \mathbb{R}$ with the *smallest* integrated 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 integrated 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 measure-informed integral,

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

$$= -2\, \mathbb{I}_\mu \left[\iota \right] + 2\, r_0,$$

to zero,

$$0 = \left. \frac{\mathrm{d}}{\mathrm{d}r_0} \mathbb{I}_\mu \left[d_{r_0}^2 \right] \right|_{r_0 = r_0^*}$$

$$0 = -2\, \mathbb{I}_\mu \left[\iota \right] + 2\, r_0^*$$

$$2\, r_0^* = 2\, \mathbb{I}_\mu \left[\iota \right]$$

$$r_0^* = \mathbb{I}_\mu \left[\iota \right].$$

In other words the measure-informed integral of the embedding function $\mathbb{I}_\mu[\iota]$ 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 $\mathbb{I}_\mu[\iota]$ 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 integrated squared distance from the mean

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

quantifies how *strongly* μ concentrates around its centrality; the larger the measure-informed integral the more diffuse the concentration is. This is known as the **variance** of μ.
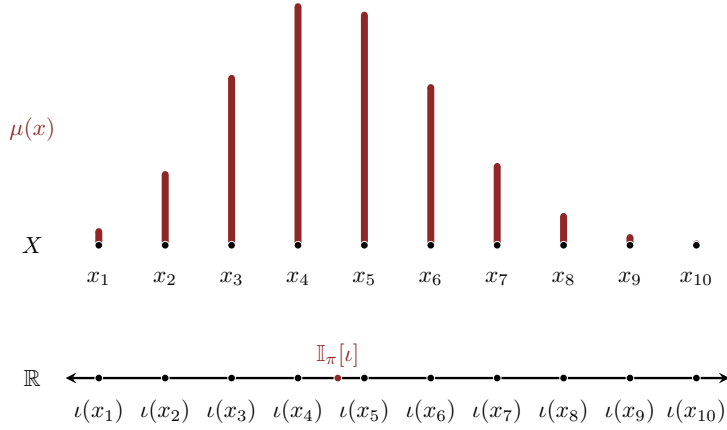
Figure 8: The measure-informed integral of an embedding function, $\mathbb{I}_\mu[\iota]$, is a continuous value even when the ambient space is discrete. In these cases the centrality of a measure can fall "between" the individual elements.

Higher-powers extract even more information. For example the measure-informed integral of the cubic integrand

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

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

The many measure-informed integral that we can construct from an embedding function can be systemized in various ways. For example the direct powers

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

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

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

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

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

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 three different ways. Firstly an isometric injection from the ambient space into a real line might not exist. Secondly if an isometric injection does exist then it needs to be not only $\mathcal{X}$-measurable but also μ-integrable. In practice measurability is almost never an issue on finite dimensional ambient spaces, but we do need to take care to check existence and integrability to avoid nonsensical results.

Consider, for example, a circular ambient space $X = \mathbb{S}^1$ equipped with a metric that assigns distances based on the angle spanned by any two points. We can construct an injective real-valued function $f : \mathbb{S}^1 \to \mathbb{R}$ by first cutting the circle at any point, unrolling it into the half-open interval $(0, 2\pi]$, and then mapping the half-open interval into the entire real line.

Unfortunately any function constructed this way will not be isometric; two points $x_1, x_2 \in \mathbb{S}^1$ around the cut will be close to each other in the circle but are mapped into points $y_1, y_2 \in (0, 2\pi]$, and then $z_1, z_2 \in \mathbb{R}$, that are far apart from each other (Figure 9).

Ultimately one can use the topological incompatibility between the circle and the real line that we first encountered in Chapter Two to show that there is no way to construct *any* isometry from the circle $\mathbb{S}^1$ into a real line $\mathbb{R}$, let alone an injective one.

This formal definition of moments is easy to dismiss as overly technical. Unfortunately the practical consequences are critical when working with spaces like circles, spheres, torii, and more. Many analyses on these spaces have been undermined by attempts to summarize measures with moments that don't actually exist!

All of this said we still to take care with the necessary conditions when working with more familiar spaces as well. For example in Section 5.2.2 we'll learn that the identify function from a real line into itself is not integrable with respect to the Lebesgue measure on a real line. Consequently the Lebesgue measure does not have a mean, let alone a variance or other higher-order moments.
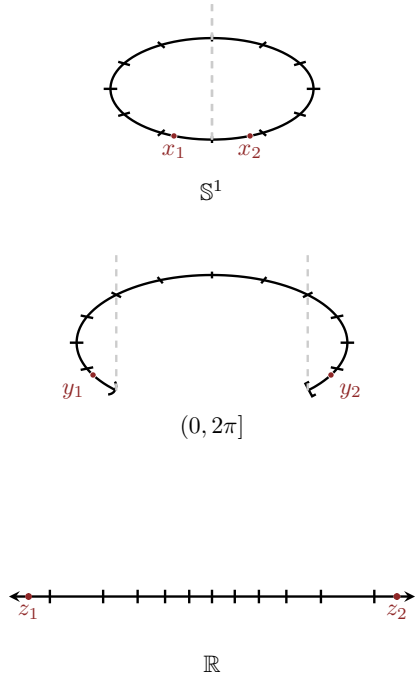
Figure 9: A circle $\mathbb{S}^1$ can be injectively mapped into a real line $\mathbb{R}$ by first cutting the circle into a half-open interval $(\theta_0 - \pi, \theta_0 + \pi]$ and then mapping the half-open interval into a real line. This mapping, however, is not isometric. Two points $x_1, x_2 \in \mathbb{S}^1$ around the cut will be close to each other in the ambient space but map into points $y_1, y_2 \in (0, 2\pi]$, and then $z_1, z_2 \in \mathbb{R}$, that are far from each other. Without an isometric injection we cannot define moments for any measure over a circle.

4.2 Histograms

When the structure of the ambient space distinguishes certain subsets the corresponding indicator functions become natural integrands to consider. Conveniently the measure-informed integrals 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

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

We can then use disjoint intervals to study the behavior of a measure by investigating how the measure allocations, or equivalently the measure-informed integrals 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 the interval (x_1, x_{B+1}) 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 measure-informed integral of the indicator function corresponding to each these sub-intervals gives the allocated measure,

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

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 10). 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 14). 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 measure-informed integrals we have to compute in order to construct the histogram (Figure 12). In practice we
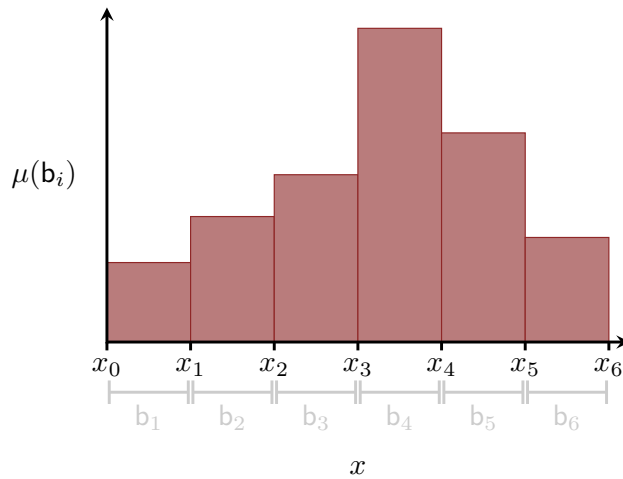
Figure 10: A histogram allows us to visualize the behavior of a measure over an ordered space. After partitioning a segment of the ambient space into disjoint intervals, or bins, the measure allocated to each bin is represented by a rectangle.

have to 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 x_1 and above x_{B+1} 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 to where the allocations start to become negligible.

On discrete measure 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 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,

$$\mathsf{I}_x = \{x' \in X \mid x' \leq x\}.$$
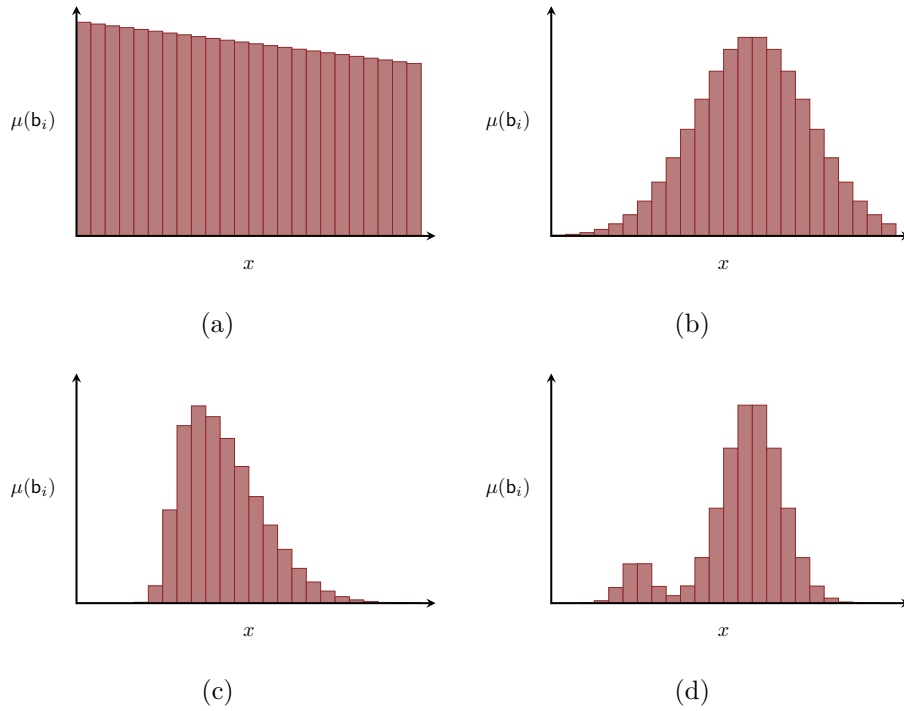
Figure 11: Histogram are extremely effective at communicating the basic features of a measure. The measure in (a) is diffuse but decaying, allocating more measure at smaller points than larger points. Conversely the measure in (b) concentrates around a single point while the measure in (c) concentrates 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).
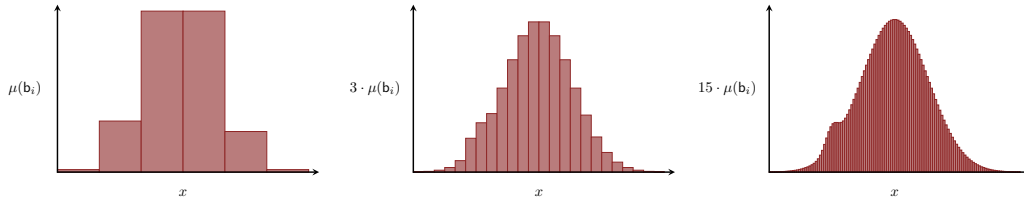
Figure 12: A histogram with a finer binning communicates more detail about a given measure, but also requires the computation of more measure-informed integrals and hence is more expensive to construct. Here as we use smaller bins we start to resolve a small side mode. Note that as we decrease the bins 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.

27

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

$$M : X \to [0, \mu(X)]$$
$$x \mapsto M_\mu(x) = \mu(\mathsf{I}_x) = \mathbb{I}_\mu[I_{\mathsf{I}_x}].$$

According this mapping is known as a **cumulative distribution function** (Figure 13).
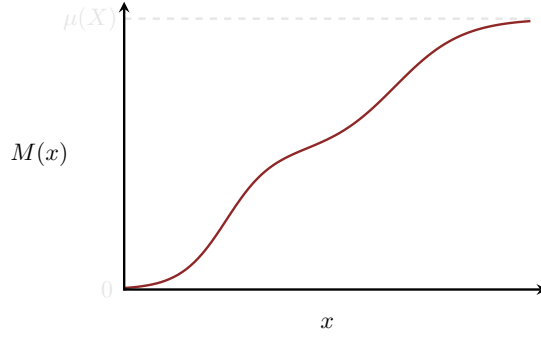
Figure 13: A cumulative distribution function quantifies how measure is allocated to expanding intervals on an ordered space. At the lower boundary of the space the interval contains no points and the cumulative distribution function returns zero. As we move towards larger values the interval expands, 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 as $\mu([x' < x])$ 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 $\mathsf{I}_{x_1} \subset \mathsf{I}_{x_2}$. Consequently

$$\mu(\mathsf{I}_{x_1}) \le \mu(\mathsf{I}_{x_2})$$

or, equivalently,

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

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 14a). In general the faster the cumulative distribution function increases the

stronger the concentration will be (Figure 14b). 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 14c).
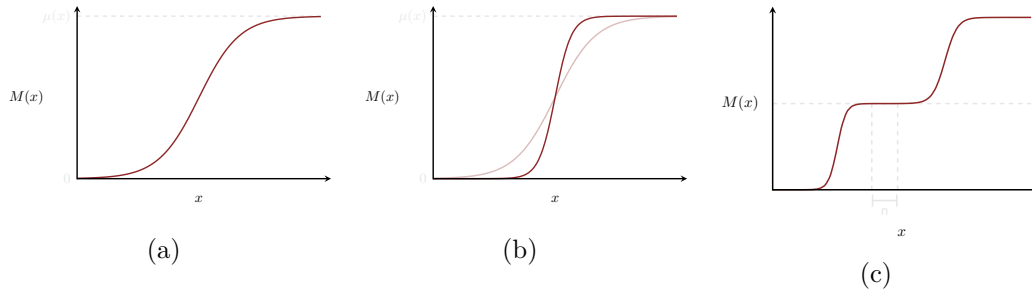

(a)   (b)

(c)

Figure 14: 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 n 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

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

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

$$\mathsf{I}_{x_2} = \mathsf{I}_{x_1} \cup (x_1, x_2].$$

Because measure allocations are additive this implies that

$$\mu(\mathsf{I}_{x_2}) = \mu(\mathsf{I}_{x_1} \cup (x_1, x_2])$$
$$= \mu(\mathsf{I}_{x_1}) + \mu((x_1, x_2])$$

or

$$\mu(\mathsf{I}_{x_2}) = \mu(\mathsf{I}_{x_1}) + \mu((x_1, x_2])$$
$$M(x_2) = M(x_1) + \mu((x_1, x_2])$$
$$\mu((x_1, x_2]) = M(x_2) - M(x_1).$$

In words the measure allocated to any half-open interval can be computed by subtracting the cumulative distribution function outputs at the interval boundaries (Figure 15).

29

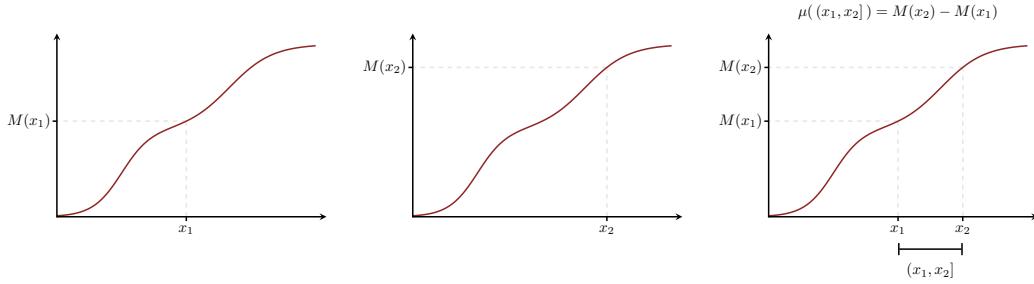$$\mu(\,(x_1, x_2]\,) = M(x_2) - M(x_1)$$

Figure 15: The difference of cumulative distribution function outputs at two points is equal to the measure allocated to the half-open 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 it 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, discrete measure 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 16).
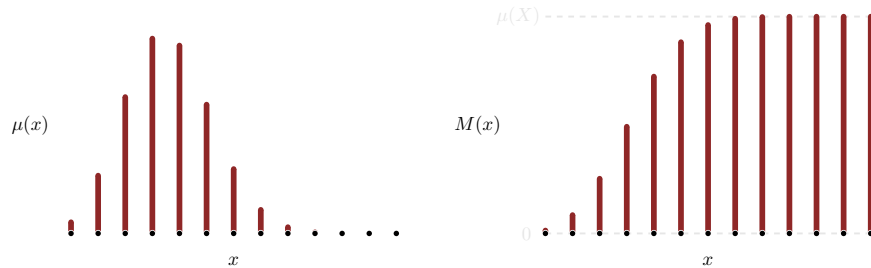
Figure 16: On ordered, discrete measure spaces a mass function and cumulative distribution function provide equivalent, and hence redundant, characterizations of a measure.

30

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 lack. 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 17).
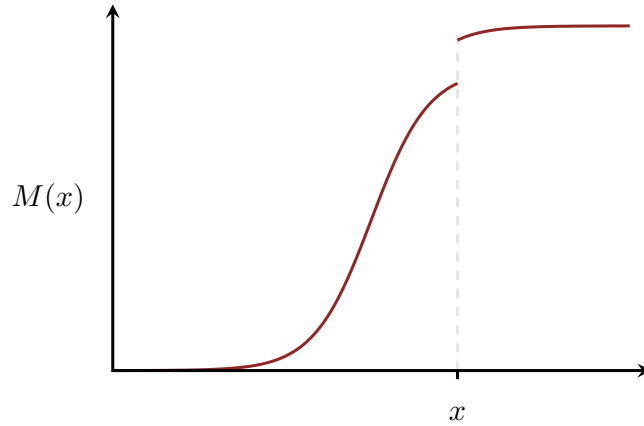
Figure 17: When the ambient space is ordered but uncountable and every atomic subset is a null subset then the cumulative distribution function will be continuous. Any discontinuities in a cumulative distribution function correspond to exceptional atomic subsets that have been allocated finite measure.

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

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

we can invert it to map any accumulated measure to the point at which that accumulation is achieved (Figure 18),

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

This inverse mapping is known as a **quantile function**.
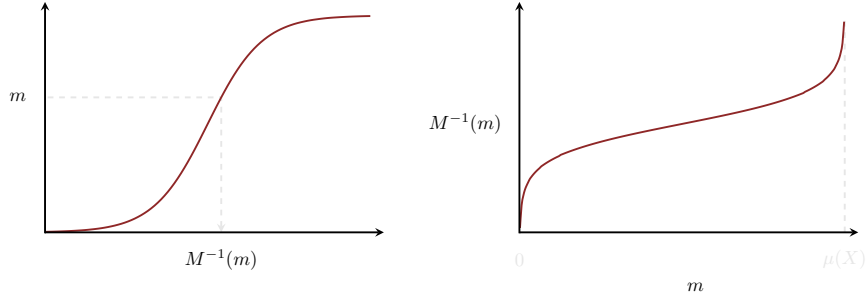
Figure 18: If a cumulative distribution function is invertible then its inverse defines a quantile function that maps accumulated measures to the points in the ambient space where the accumulation is reached.

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,

$$\Pi(x_{0.5}) = 0.5,$$

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,

$$\Pi(x_{0.25}) = 0.25$$
$$\Pi(x_{0.75}) = 0.75,$$

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,

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

and bounded above by the point x_+ that achieves the smallest accumulated measure above m (Figure 19),

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

Many software packages implement *heuristic* quantile functions that either return x_{m-} or x_{m+} or interpolate between x_{m-} and x_{m+} to provide a single value when the cumulative distribution function is not invertible. In this case different interpolation strategies define different quantile functions.
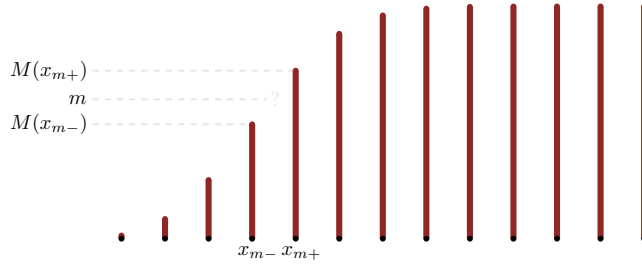
32

Figure 19: 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.

5 Explicit Measure-Informed Integrals

Up to this point our discussion of measure-informed integration has been theoretical. Given a measure μ we have shown that a linear map from sufficiently nice, real-valued functions f to real numbers $\mathbb{I}_\mu[f] \in \mathbb{R}$ is well-defined. We do not yet know, however, how to *evaluate* that map to give explicit measure-informed integrals in practice.

Fortunately a few exceptional measures produce integrals that can be computed from certain explicit mathematical operations that allow us to realize them in practice. In this section we'll review these exceptional measures and their practical consequences. Along the way we'll also see how measure-informed integration relates to the Riemann integral from calculus.

5.1 Integration on Discrete Measure Spaces

Because they can be completely specified by a mass function, measure allocations on discrete measure spaces are particularly straightforward to implement in practice. Conveniently measure-informed integrals on these spaces are also completely specified by mass functions.

5.1.1 Integration As Summation

For any discrete 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 integral with respect to any measure μ follows immediately by applying linearity,

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

and then the definition of measure-informed integrals for indicator functions,

$$\mathbb{I}_\mu[f] = \sum_{x' \in X} f(x') \cdot \mathbb{I}_\mu \left[I_{\{x'\}} \right]$$
$$= \sum_{x' \in X} f(x') \cdot \mu(\{x'\}).$$

Consequently measure-informed integrals with respect to *any* real-valued function on discrete measure spaces reduces to summations which we can compute explicitly. Moreover because the summations are informed by only the measure allocations to atomic subsets they can be computed using only the mass function and not the entire measure.

If X is not only countable but also finite then the general definition of measure-informed integral reduces to the heuristic construction that we considered in Section 1.

5.1.2 Practical Consequences

When X is finite we can implement the summation given by any measure-informed integral by directly looping over each integrand 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 measure-informed integral exactly. That said we may be able to *approximate* them by summing over only a finite number of elements where both $\mu(x)$ and $f(x)$ are large. The more terms we include, the better these finite sums will approximate the exact measure-informed integrals.

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 integral 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,

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

In other words all integrals with respect to the counting measure can be implemented by simply summing over the integrand 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 integrals

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

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 start, the machinery of measure-informed integration allows us to prove that this intuitive definition is consistent with how measure theory behaves more generally.

5.2 Integration on Real Lines

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

5.2.1 Lebesgue Verses Riemann

By definition the measure-informed integral of an indicator function is given by the measure allocated to the defining subset,

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

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 integral of an interval indicator function is given by

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

That integral, however, also happens to be the area under the curve defined by the corresponding indicator function (Figure 20a),

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

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

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

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,

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

36

By linearity, however, this is just the measure-informed integral of the simple function itself (Figure 20b)

$$
\begin{aligned}
\text{area} &= \sum_j \mathbb{I}_\lambda[\phi_j \cdot I_{[x_{1,j}, x_{2,j}]}] \\
&= \mathbb{I}_\lambda[\sum_j \phi_j \cdot I_{[x_{1,j}, x_{2,j}]}] \\
&= \mathbb{I}_\lambda[s].
\end{aligned}
$$
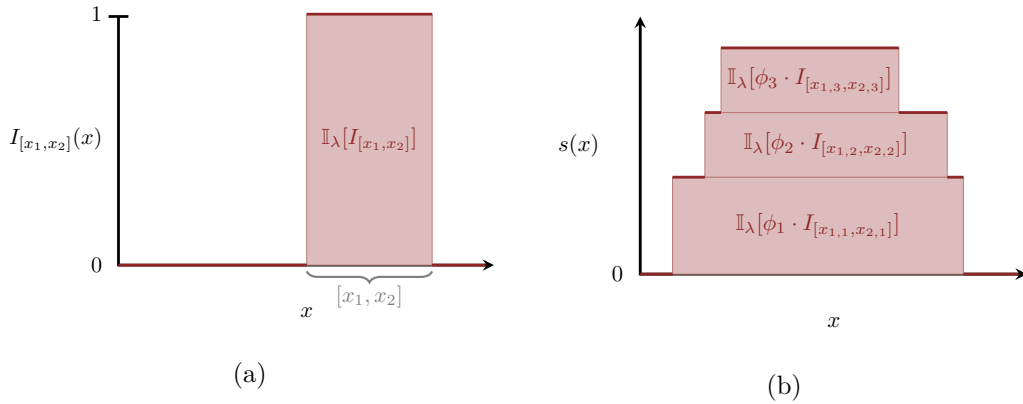
Figure 20: Integrals of simple functions with respect to the Lebesgue measure are intimately related to the area under the curve defined by simple functions. (a) The area under the curve defined by an interval indicator function is equal to the height, 1, times the length of the interval. That, however, is just equal to the Lebesgue integral of the indicator function itself. (b) The area under the curve defined by interval simple functions is built up from the area of rectangles defined by each component indicator function. The total area is equal to the Lebesgue integral 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 Lebesgue integrals, and on the other we can use it to compute the area under the curve defined by any sufficiently nice function (Figure 21a).

We can also use classic 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 length δ becomes smaller and smaller the sum of the rectangle areas converges to the area under the curve (Figure 21b),

$$
\int \mathrm{d}x \, f(x) = \lim_{\delta \to 0} \sum_{n=-\infty}^{\infty} \delta \cdot f(x_0 + n \cdot \delta).
$$
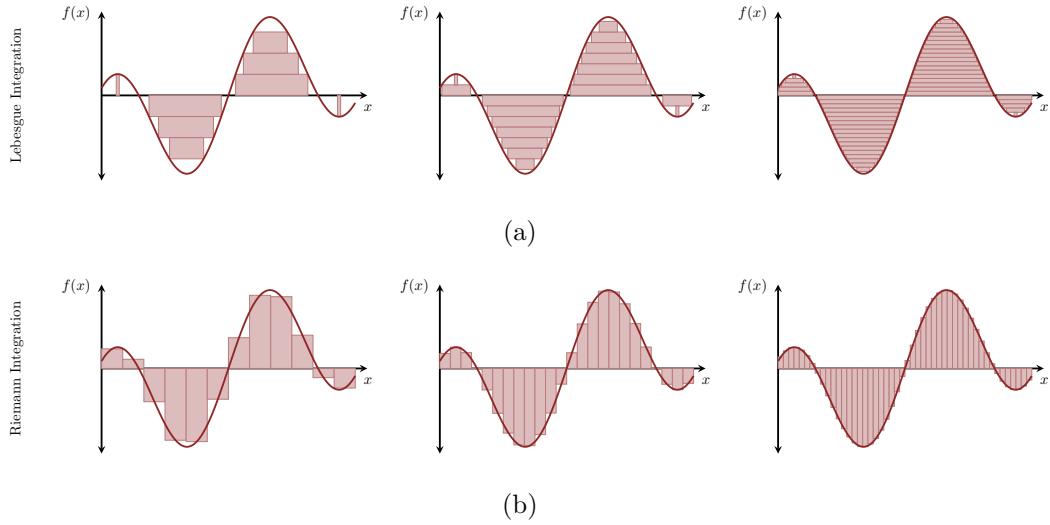
Figure 21: On a real line $X = \mathbb{R}$ integration with respect to the Lebesgue measure and Riemann integration both quantify the area under a curve defined by a sufficiently nice real-valued function $f : \mathbb{R} \to \mathbb{R}$. (a) As we add more components the Lebesgue integral of a simple function converges to the Lebesgue integral of f. At the same time the sum of the rectangular areas defined by each component indicator function converges to the area under the curve defined by f. (b) Riemann integration computes the area under the curve as a sum of increasingly narrow rectangular areas, only the rectangles are stacked horizontally instead of vertically.

Geometrically Lebesgue integration computes 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 integrals with respect to the Lebesgue measure to classic integration!

More formally for any measurable and λ-integrable real-valued function $f : \mathbb{R} \to \mathbb{R}$ we have

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

so long as the Riemann integral $\int \mathrm{d}x\, f(x)$ is well-defined. In other words Lebesgue integration on the real line completely generalizes Riemann integration; measure-informed integration then generalizes Lebesgue integration on the real line to arbitrary meaures and spaces.

This particular equivalence is the motivation for the many alternative notations that we discussed in . In general the integral signs in those notations do not correspond to the Riemann integral of calculus, 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 Riemann integral then we can apply the tools of calculus to evaluate Lebesgue integrals. The exceptional Riemann integrals that can be evaluated analytically allow us to compute the corresponding Lebesgue integrals exactly. More generally we can use to numerical integration techniques to approximate the Riemann integrals, and hence approximately evaluate Lebesgue integrals.

For example the measure-informed integral of an interval indicator function is given by

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

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 Riemann integral. In order to properly relate Lebesgue integrals to Riemann integrals we have to fix the *orientation* of the intervals.

Similarly the mean of a Lebesgue measure would by given by the integral of the identity function,

$$
\begin{aligned}
\mathbb{I}_\lambda[\iota] &= \int_{-\infty}^{\infty} \mathrm{d}x\, \iota(x) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x\, x \\
&= \left.\frac{1}{2}x^2\right|_{-\infty}^{\infty} \\
&= \infty - \infty.
\end{aligned}
$$

Unfortunately this result is ill-posed 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,

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

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

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

In particular the measure allocated to any interval becomes

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

By appropriately scaling the Lebesgue measure in this way we can implement all kinds of measures over a real line, including most probability distributions of practical interest. We'll formalize this procedure in the next chapter.

6 Conclusion

Measure-informed integrals are the main way that we interact with measures, both in theory and in practice. Equivalently expectation values are the main way that we can probe the behavior of probability distributions. 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 extend the exceptionally explicit integrals with respect to Lebesgue measures to a much larger class of measures, including many probability distributions. Later on we'll learn some powerful *sampling* techniques for directly estimating expectation values for general probability distributions.

Acknowledgements

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

A repository containing all of the files used to generate this chapter is available on GitHub.