These Notes Are Works-In-Progress

A Measure-Theoretic Approach to Bayesian Statistics

or, Bayesian Statistics for Mathematicians

Raymond T. Melton

July 14, 2023

Contents

Preface

THESE notes were never initially intended for anyone else to look at. I am working to make them easier on the eyes.

The Hilbert space stuff is off a bit, I know. It was typed up long before I worked through the details of quotient spaces, like L_p, so there are a few mis-statements yet to be corrected. Actually, it is mathematically correct in the sense that separable Hilbert space people tend to use '=' when they mean 'isomorphic.' Still.

Feel free to contact me. I am willing to share ideas.

1 Introduction

Many readers will have noticed the recent trend toward quotations at the beginnings of chapters in scientific books. Often these quotes are sappy, dippy little things as if the authors of the book were struggling for profundity.

Heterochrony: The Evolution of Ontogeny
— M. L. McKinney

The real reason these notes exist is so I could remember what progress I had made. As a mathematician, I was trying to make headway into statistics. I am putting these notes online, thinking they may help another mathematician trying to do the same.

What is the difference between these notes and other texts? We will not make topological assumptions when it comes to regular conditional distributions. We will not confuse classes in $L_1(X, \mathfrak{A}, \mu)$ with functions in $\mathscr{L}_1(X, \mathfrak{A}, \mu)$. We will not expect the reader to "keep this, that, and the other thing in mind." We spell it out instead. But the main difference is that the measure theory here is based

upon semi-rings as opposed to σ-algebras, to be explained further in the next few paragraphs.

These notes exist in part because of the way I had learned measure theory. After learning measure theory, I was trying to learn statistics, and I mean the level of statistics that begins with the Radon-Nikodym theorem, and Kolmogorov's conditional mathematical expectation, and goes from there. In all the statistics books that I looked, the Lebesgue integral was defined in terms of a measure on a σ-algebra. But the measure theory I had learned was based upon a measure on a semi-ring, with the corresponding Lebesgue integral defined in terms of the restriction of a certain outer measure, namely the Carathéodory extension. What is the difference between the theory that results from each of the two approaches?

One main difference: In the case of a measure on a σ-algebra and the consequent Lebesgue integral, say as developed in Loeve [5], the resulting conditional mathematical expectation is a class of functions, where each function in the class is measurable with respect to a σ-subalgebra, but in the case of a measure on a semi-ring, as developed in Aliprantis & Burkinshaw [2], the consequent conditional mathematical expectation is a class of functions from which we can select a representative function which is measurable with respect to the σ-subalgebra. Not a big difference, but you still need to make the selection. Then again, you always need to make a selection with a regular conditional probability. Right?

These notes also exist in part because I wanted to see clearly which σ-algebras were involved in the statistical models. Making the domain and codomain of each function clear is simply a matter of course for any real mathematician. There are expositions of probability theory where random variables are considered primitive in the sense that no mention of the domain of such measurable

functions are thought needed. This is not one of those kinds of exposition. The struggle to have a probability theory where random variables are primitive reaches back through time immemorial. Topological assumptions in probability theory are always the sign of a struggle. Like the struggle a mathematician might put themselves through, mistakenly thinking that they always have to be working in a category.

Initially, when trying to learn statistics, I had been trying to understand regression models, and had a background in functional analysis and operator theory, but little in the way of statistics. The hopeful outcomes at the start were:

1. Understand Kolmogorov's *conditional mathematical expectation.*
2. Explain regression models to myself in a clear way.
3. Make real mathematical sense out of the bias-variance tradeoff.

It turns out that conditional mathematical expectation is an application of the Radon-Nikodym theorem. Regression models are nothing but projections. And there is a decomposition of conditional expectation into a conditional variance and a squared conditional bias because conditional mathematical expectation is a so-called averaging operator. Not that I was disappointed.

Statistical models only started making sense after learning Bayesian statistical models. That is all you will find here. All of the material regarding frequentist statistics has been edited out, like when a cocoon is left behind.

The notes that follow might be described as more of a measure-theoretic companion to something comprehensive and explanatory on applied Bayesian statistical models. I strongly recommend Richard McElreath's *Statistical Rethinking* [6].

Decisions to be made over notation favor cumbersome-but-clear over cheap or ambiguous. The proofs of some of the propositions are wildly over-detailed, especially on paths which are thought to be critical. Material comes from multiple sources, some of which might be listed in the references.

One more thing. A quote from *Thinking Like Your Editor,* by Rabiner & Fortunato, 2002, where the question is raised, and I really hope I am paraphrasing its content correctly here: Once you've defined your audience, to which level of sophistication do you write?

> The answer is that if you want to write serious nonfiction, and if you want your work to be taken seriously, you never dumb down. You always write up. Writing down runs the risk of having your book read like a young adult book. Assume a high reader level of intellectual sophistication, even though you suspect their knowledge of your particular subject may not be high. Fully explain, but do not simplify.

This is not a children's reader. No attempt has been made to dumb down anything.

2 Measure Spaces and Measurable Sets

THIS is only an abbreviated review of measure theory, based upon semirings. Roughly, a measure on a semiring is extended to the powerset, and then restricted to the measurable subsets. Resulting measure spaces are complete in the sense that every subset of a set of measure zero is Carathéodory measurable. That is, every null set is measurable. It remains to be seen whether this particular notion of completeness will be sufficient. (Sorry about the pun.)

Any missing detail might be found in Aliprantis and Burkinshaw, *Principles of Real Analysis* [2]. Except Proposition 2.23. That was found in Aliprantis and Border's book, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Third Edition, 2006; see 10.31 Lemma therein, page 386.

Definition 2.1. Let X be a nonempty set. A collection $\mathfrak{S}$ of subsets of X is a **semiring** if it has the following three properties:

1. $\varnothing \in \mathfrak{S}$.
2. If A and B are in $\mathfrak{S}$, then $A \cap B \in \mathfrak{S}$.
3. If A and B in $\mathfrak{S}$, there exist disjoint sets $C_1, \ldots, C_n$ in $\mathfrak{S}$ such that $A \setminus B = \bigcup_{i=1}^{n} C_i$.

Definition 2.2. Let $\mathfrak{S}$ be a semiring of subsets of a set X. A set function $\mu : \mathfrak{S} \to [0, \infty]$ is a **measure** if it has both of the following

properties:

1. $\mu(\varnothing) = 0$.
2. μ is σ-additive.

Definition 2.3. A triple $(X, \mathfrak{S}, \mu)$, where X is a nonempty set, the collection $\mathfrak{S}$ is a semiring of subsets of X, and μ is a measure, is called a **measure space**.

Definition 2.4. A nonempty collection $\mathfrak{A}$ of subsets of a nonempty set X is an **algebra** if it is closed with respect to finite intersections and complements; that is, if it has both the following properties:

1. If A and B are in $\mathfrak{A}$, then $A \cap B \in \mathfrak{A}$.
2. If $A \in \mathfrak{A}$, then $A^{\mathsf{c}} \in \mathfrak{A}$.

Consequently, both $\varnothing \in \mathfrak{A}$ and $X \in \mathfrak{A}$.

> A concept in between those of semirings and algebras of sets is that of rings of sets. A nonempty collection of subsets of a set is a **ring** if it is closed with respect to set differences and finite unions. Every algebra of sets is also a ring of sets. Every ring of sets is also a semiring of sets.

Definition 2.5. An algebra of subsets of a set is a σ**-algebra** if it is closed with respect to countable unions, in which case it must also be closed with respect to countable intersections. If $\mathfrak{A}$ is a σ-algebra and $\mathfrak{B} \subseteq \mathfrak{A}$, then $\mathfrak{B}$ is a σ**-subalgebra** of $\mathfrak{A}$ if $\mathfrak{B}$ is a σ-algebra. If $\mathfrak{F}$ is a collection of subsets of a nonempty set, then $\mathfrak{F}$ is included in a smallest σ-algebra, denoted $\sigma(\mathfrak{F})$. It is the intersection of all σ-algebras containing $\mathfrak{F}$, and is called the σ**-algebra generated** by $\mathfrak{F}$.

Definition 2.6. Let $\mathfrak{P}(X)$ be the power set of a set X. A set function $\mu : \mathfrak{P}(X) \to [0, \infty]$ is an **outer measure** if it has the following three properties:

1. $\mu(\varnothing) = 0$.
2. μ is monotone. That is, if A and B are in $\mathfrak{P}(X)$ with $A \subseteq B$, then $\mu(A) \leq \mu(B)$.
3. μ is σ-subadditive. That is, $\mu\big(\bigcup_{n=1}^{\infty} A_n\big) \leq \sum_{n=1}^{\infty} \mu(A_n)$ for every sequence $\{A_n\}$ of subsets of X.

Definition 2.7 (Carathéodory). Let μ be an outer measure on the powerset $\mathfrak{P}(X)$ of some set X. A subset B of X is a μ**-measurable set** if

$$\mu(A) = \mu(A \cap B) + \mu(A \cap B^{\mathsf{C}})$$

for all $A \subseteq X$. Let μ-$\mathfrak{Meas}$ denote the collection of all μ-measurable subsets of X.

Remark 2.8. Measurable catch phrases:

- A set is μ-measurable if it sharply cuts every set.
- A set is μ-measurable if you can always make change with it.

Definition 2.9. Given a measure space $(X, \mathfrak{S}, \mu)$, define a set function $\mu^* : \mathfrak{P}(X) \to [0, \infty]$ by

$$\mu^*(A) = \inf\left\{\sum_{n=1}^{\infty} \mu(A_n) : \{A_n\} \text{ is a sequence of } \mathfrak{S} \text{ with } A \subseteq \bigcup_{n=1}^{\infty} A_n\right\},$$

where $\inf \varnothing = \infty$. The set function μ^* is the **outer measure generated** by μ.

It will turn out that μ^*, when restricted to the collection of μ^*-measurable sets, is an honest-to-God measure. We might also denote this restriction by μ^* for now. If this reduces clarity at any point, then we will introduce some fairly cumbersome notation, like $\mu_{|\mu\text{-}\mathfrak{Meas}}$, or $\mu^*{}_{|\mu^*\text{-}\mathfrak{Meas}}$.

Proposition 2.10. *The outer measure generated by a measure really is an outer measure.*

Proposition 2.11. *Let $(X, \mathfrak{S}, \mu)$ be a measure space. The outer measure μ^* generated by μ is an extension of μ from $\mathfrak{S}$ to the powerset $\mathfrak{P}(X)$ of X. That is, if $A \in \mathfrak{S}$, then $\mu^*(A) = \mu(A)$. Equivalently, the restriction $\mu^*{}_{|\mathfrak{S}}$ of μ^* to the semiring $\mathfrak{S}$ is equal to μ.*

Notation 2.12. Let μ be an outer measure on the powerset $\mathfrak{P}(X)$ of some set X. A subset A of X is called a μ-**null set** if $\mu(A) = 0$. Any definition or relation on X that holds except possibly on a μ-null set is said to hold μ-**almost everywhere**, or simply μ-a.e.

Proposition 2.13. *Let μ be an outer measure on the powerset $\mathfrak{P}(X)$ of some set X. Every μ-null set is μ-measurable. Consequently, since μ is monotone, every subset of a μ-null set is μ-measurable.*

Proposition 2.14. *If μ is an outer measure on the powerset $\mathfrak{P}(X)$ of a set X, then the collection μ-$\mathfrak{Meas}$ of all μ-measurable subsets of X is a σ-algebra, and the triple $(X, \mu\text{-}\mathfrak{Meas}, \mu_{|\mu\text{-}\mathfrak{Meas}})$ is a measure space. That is, the outer measure μ, when restricted to the σ-algebra μ-$\mathfrak{Meas}$ of all μ-measurable subsets of X, is σ-additive.*

Corollary 2.15. *Let $(X, \mathfrak{S}, \mu)$ be a measure space. Let A and B be μ^*-measurable sets with $B \subseteq A$ and with $\mu^*(A) < \infty$. Then $\mu^*(A \cap B^{\mathsf{c}}) = \mu^*(A) - \mu^*(B)$.*

Proposition 2.16. *If $(X, \mathfrak{S}, \mu)$ is a measure space, then every element of the semiring $\mathfrak{S}$ is a μ^*-measurable subset. That is, $\mathfrak{S} \subseteq \mu^*\text{-}\mathfrak{Meas}$.*

Definition 2.17. A measure space $(X, \mathfrak{S}, \mu)$ is **finite** if $\mu^*(X) < \infty$. We may also say that the measure μ is finite. A measure space $(X, \mathfrak{S}, \mu)$ is σ-**finite** if there is a disjoint sequence $\{X_n\}$ of subsets of $\mathfrak{S}$ with $\mu(X_n) < \infty$ and $X \subseteq \bigcup_n X_n$. We may also say that measure μ is σ-finite.

Proposition 2.18. *Let $(X, \mathfrak{S}, \mu)$ be a finite measure space. A subset $E \subseteq X$ is μ^*-measurable if and only if*

$$\mu^*(X) = \mu^*(E) + \mu^*(E^\mathsf{C}).$$

Proposition 2.19. *Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $E \in \mu^*\text{-}\mathfrak{Meas}$. Define a collection of sets, the restriction of $\mathfrak{S}$ to E, by*

$$\mathfrak{S}_E = \{E \cap A : A \in \mathfrak{S}\},$$

and define a function $\nu : \mathfrak{S}_E \to [0, \infty]$ by $\nu(E \cap A) = \mu^*(E \cap A)$. Then the triple $(X, \mathfrak{S}_E, \nu)$ is a measure space, and $\nu^* = \mu^*_{|\mathfrak{P}(E)}$. Furthermore, $\nu^*\text{-}\mathfrak{Meas} = \{F \subseteq E : F \in \mu^*\text{-}\mathfrak{Meas}\}$.

Proposition 2.20. *Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $A \subseteq X$. If there is an $E \in \mu^*\text{-}\mathfrak{Meas}$ with $A \subseteq E$ such that $\mu^*(E) < \infty$ and $\mu^*(E) = \mu^*(A) + \mu^*(E \cap A^\mathsf{C})$, then $A \in \mu^*\text{-}\mathfrak{Meas}$.*

Proposition 2.21. *Let $(X, \mathfrak{S}, \mu)$ be a σ-finite measure space, let $\mathfrak{T}$ be a semiring of subsets of X with $\mathfrak{S} \subseteq \mathfrak{T} \subseteq \mu^*\text{-}\mathfrak{Meas}$, and let $\nu : \mathfrak{T} :\to [0, \infty]$ be a measure on $\mathfrak{T}$. If $\nu = \mu$ on $\mathfrak{S}$, then $\nu = \mu^*$ on $\mathfrak{T}$. Consequently, the measure $\mu^* : \mu^*\text{-}\mathfrak{Meas} \to [0, \infty]$ is the unique extension of the measure μ from the semiring $\mathfrak{S}$ to the σ-algebra $\mu^*\text{-}\mathfrak{Meas}$.*

Proposition 2.22. *Let $(X, \mathfrak{S}, \mu)$ be a measure space. If $A \subseteq X$, then there is a superset $B \in \sigma(\mathfrak{S})$ of A with $\mu^*(A) = \mu^*(B)$.*

Proof. Let $A \subseteq X$. We first look at the case where the outer measure of A is infinite; let $\mu^*(A) = \infty$. We can take $B = X$; we know $X \in \sigma(\mathfrak{S})$ because $\varnothing \in \mathfrak{S}$ and algebras (therefore σ-algebras) are closed with respect to complements. And outer measures are monotone, so $\mu^*(A) \leq \mu^*(X)$. But this implies that $\mu^*(X) = \infty$, so $\mu^*(A) = \mu^*(X)$, completing the proof of this case.

We now look at the complementary case; let $\mu^*(A) < \infty$. For each $k \in \{1, 2, \ldots\}$ there must be, according to the definition of $\mu^*(A)$, a sequence $\{A_1^k, A_2^k, \ldots\}$ of $\mathfrak{S}$ whose union covers A, where

$$\mu^*(A) \leq \sum_{n=1}^{\infty} \mu(A_n^k) < \mu^*(A) + 1/k.$$

In this case, we can take $B = \bigcap_{k=1}^{\infty} \big(\bigcup_{n=1} A_n^k\big)$, which is in $\sigma(\mathfrak{S})$. Because $A \subseteq \bigcup_{n=1} A_n^k$ for each k, it follows that $A \subseteq B$. Also,

$$\mu^*(A) \leq \mu^*(B) \leq \mu^* \left(\bigcup_{n=1} A_n^k \right) \leq \sum_{n=1} \mu^*\big(A_n^k\big) = \sum_{n=1} \mu\big(A_n^k\big) < \mu^*(A) + 1/k.$$

Since this holds for each $k \in \{1, 2, \ldots\}$, it follows that $\mu^*(A) = \mu^*(B)$, as required. ■

Proposition 2.23. *Let $(X, \mathfrak{S}, \mu)$ be a σ-finite measure space. If $A \in \mu^*$-$\mathfrak{Meas}$, then there is a subset $C \subseteq X$ with $A \cap C = \varnothing$ and such that $\mu^*(C) = 0$ with $A \cup C \in \sigma(\mathfrak{S})$.*

Proof. Let $A \in \mu^*$-$\mathfrak{Meas}$. The measure space $(X, \mathfrak{S}, \mu)$ is supposed to be σ-finite, so there is a sequence $\{X_n\}$ of subsets of $\mathfrak{S}$ with $\mu(X_n) < \infty$ and $X \subseteq \bigcup_n X_n$. By Proposition 2.22, for each subset $A \cap X_n$ there is a subset $B_n \in \sigma(\mathfrak{S})$ with $A \cap X_n \subseteq B_n$ and $\mu^*(A \cap X_n) = \mu^*(B_n)$. We take $C = \big(\bigcup_n B_n\big) \cap A^{\mathsf{c}}$ and show that C has the required properties.

Pure set theory shows that $C \subseteq \bigcup_n \big(B_n \cap (A \cap X_n)^{\mathsf{c}}\big)$. Since $\mu(X_n) = \mu^*(X_n) < \infty$, it follows that $\mu^*(A \cap X_n) = \mu^*(B_n) < \infty$ and so we can apply Corollary 2.15 which says $\mu^*(B_n \cap (A \cap X_n)^{\mathsf{c}}) = \mu^*(B_n)\mu^*(A \cap X_n)$. But $\mu^*(B_n) = \mu^*(A \cap X_n)$. This means $\mu^*(B_n \cap (A \cap X_n)^{\mathsf{c}}) = 0$ which implies that $\mu^*\big(\bigcup_n \big(B_n \cap (A \cap X_n)^{\mathsf{c}}\big)\big) = 0$. By monotonicity, $\mu^*(C) = 0$.

The fact that $C \subseteq A^{\mathsf{c}}$ implies that $A \cap C = \varnothing$. Finally, to see that $A \cup C \in \sigma(\mathfrak{S})$, since we already know that each B_n is in $\sigma(\mathfrak{S})$, we simply need to show that $A \cup C = \bigcup_n B_n$. Again, this is pure set theory. Briefly, since $A \subseteq \bigcup_n (A \cap X_n) \subseteq \bigcup_n B_n$, it follows that $A \cup (\bigcup_n B_n) = \bigcup_n B_n$, and so

$$\begin{aligned} A \cup C &= A \cup \left((\bigcup_n B_n) \cap A^{\mathsf{c}} \right) \\ &= \left(A \cup \bigcup_n B_n \right) \cap (A \cup A^{\mathsf{c}}) \quad \text{just distribute} \\ &= A \cup \bigcup_n B_n \\ &= \bigcup_n B_n, \end{aligned}$$

as required. ■

3 Measurable Functions and Integration

THROUGHOUT this section, let $(X, \mathfrak{S}, \mu)$ denote a measure space, so $\mathfrak{S}$ is at least a semiring, and we continue to let μ^*-$\mathfrak{Meas}$ denote the σ-algebra of μ^*-measurable subsets of X.

Notation 3.1. Let X and Y be sets, and let $\mathfrak{P}(X)$ and $\mathfrak{P}(Y)$ denote their respective powersets. For any function $f : X \to Y$, the set X is the **domain** of f and the set Y is the **codomain**. For any subset $A \subseteq X$, the set

$$f(A) \coloneqq \{f(x) : x \in A\}$$

is the **image** of A under f, which we could denote $f^{\rightarrow}(A)$. The subset $f(X) \subseteq Y$ is the **range** of f. And for any subset $B \subseteq Y$, the set

$$f^{\leftarrow}(B) \coloneqq \{x \in X : f(x) \in B\}$$

is the **preimage** of B under f. We might just as well have said let $f^{\leftarrow}$ denote the image of f under a contravariant powerset functor. This means that

$$f^{\leftarrow} : \mathfrak{P}(Y) \to \mathfrak{P}(X) : B \mapsto f^{\leftarrow}(B).$$

The set $f^{\leftarrow}(B)$ might also be denoted $f^{-1}(B)$, but not here. Any restriction of such a $f^{\leftarrow}$ which maps some subset of $\mathfrak{P}(Y)$ to some subset of $\mathfrak{P}(X)$ may still be denoted $f^{\leftarrow}$, in which case the domain and codomain of $f^{\leftarrow}$ will be made explicit.

For any topological space X, we will let $\mathfrak{Bor}_X$ denote the σ-algebra of subsets of X generated by the topology, namely the **Borel subsets** of the topological space X.

Definition 3.2. A pair $(X, \mathfrak{A})$, where X is a set and $\mathfrak{A}$ is a σ-algebra of subsets of X, is called a **measurable space**. Better terminology might be something flavored Mackey.

Definition 3.3. Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces. A function $f : X \to Y$ is called an $(\mathfrak{A}, \mathfrak{B})$-**measurable function** if a restriction of $f^{\leftarrow}$ to $\mathfrak{B}$ is a function $\mathfrak{B} \to \mathfrak{A}$; that is, if $f^{\leftarrow}(B) \in \mathfrak{A}$ for each $B \in \mathfrak{B}$. To indicate that $f : X \to Y$ is $(\mathfrak{A}, \mathfrak{B})$-measurable, we might use the notation $f : (X, \mathfrak{A}) \to (Y, \mathfrak{B})$.

Remark 3.4. An attempt at a categorical approach to measurable spaces and measurable functions which does not take into account collections of null sets would be considered a first attempt. There is nothing wrong with first attempts. Who knows? You may end up factoring by an ideal.

Proposition 3.5. *Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be a measurable spaces. If μ is a measure on $\mathfrak{A}$, and a function $f : X \to Y$ is $(\mathfrak{A}, \mathfrak{B})$-measurable, where $f^{\leftarrow} : \mathfrak{B} \to \mathfrak{A}$, then the set function*

$$\mu \circ f^{\leftarrow} : \mathfrak{B} \to [0, \infty] : B \mapsto \mu\big(f^{\leftarrow}(B)\big)$$

defines a measure on $\mathfrak{B}$.

Definition 3.6. The measure $\mu \circ f^{\leftarrow}$ on the σ-algebra of the codomain of f in Proposition 3.5 is the **measure induced by** f.

Remark 3.7. Should two $(\mathfrak{A}, \mathfrak{B})$-measurable functions f and g be equal μ^*-almost everywhere, then the induced σ-algebras $f^{\leftarrow}(\mathfrak{B})$

and $g^{\leftarrow}(\mathfrak{B})$ need not be the same. However, the following proposition says that the induced measures $\mu^* \circ f^{\leftarrow}$ and $\mu^* \circ g^{\leftarrow}$ are the same.

Proposition 3.8. *Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $(Y, \mathfrak{B})$ be a measurable space. If the functions $f : X \to Y$ and $g : X \to Y$ are $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{B})$-measurable functions with $f = g$ μ^*-almost everywhere on X, then $\mu^*\big(A \cap f^{\leftarrow}(B)\big) = \mu^*\big(A \cap g^{\leftarrow}(B)\big)$ for all $A \in \mu^*\text{-}\mathfrak{Meas}$ and for all $B \in \mathfrak{B}$. So in particular, $\mu^*\big(f^{\leftarrow}(B)\big) = \mu^*\big(g^{\leftarrow}(B)\big)$ for all $B \in \mathfrak{B}$.*

Proof. Define $M = \{x \in X : f(x) = g(x)\}$, so that by hypothesis, $\mu^*(M^{\complement}) = 0$. Let $A \in \mu^*\text{-}\mathfrak{Meas}$ and $B \in \mathfrak{B}$. If $x \in f^{\leftarrow}(B)$ and $x \notin g^{\leftarrow}(B)$, then $f(x) \in B$ and $g(x) \in B^{\complement}$, implying $f(x) \neq g(x)$ and $x \notin M$, meaning

$$f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement} \subseteq M^{\complement}.$$

This says

$$\mu^*\big(f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement}\big) = 0$$

since $\mu^*(M^{\complement}) = 0$. Because $f^{\leftarrow}(B)$ is a disjoint union, as in

$$f^{\leftarrow}(B) = \big(f^{\leftarrow}(B) \cap g^{\leftarrow}(B)\big) \sqcup \big(f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement}\big),$$

we can also write $A \cap f^{\leftarrow}(B)$ as a disjoint union:

$$A \cap f^{\leftarrow}(B) = \Big(A \cap \big(f^{\leftarrow}(B) \cap g^{\leftarrow}(B)\big)\Big) \sqcup \Big(A \cap \big(f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement}\big)\Big).$$

Because $A \cap f^{\leftarrow}(B)$ is μ^*-measurable, it follows by the very definition of what it means to be measurable that

$$\mu^*\big(A \cap f^{\leftarrow}(B)\big) = \mu^*\big(A \cap f^{\leftarrow}(B) \cap g^{\leftarrow}(B)\big) + \mu^*\big(A \cap f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement}\big).$$

But the fact that $\mu^*\big(f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement}\big) = 0$ implies that $\mu^*\big(A \cap f^{\leftarrow}(B) \cap (g^{\leftarrow}(B))^{\complement}\big) = 0$, and so

$$\mu^*\big(A \cap f^{\leftarrow}(B)\big) = \mu^*\big(A \cap f^{\leftarrow}(B) \cap g^{\leftarrow}(B)\big).$$

By a symmetric argument,

$$\mu^*\big(A \cap g^{\leftarrow}(B)\big) = \mu^*\big(A \cap f^{\leftarrow}(B) \cap g^{\leftarrow}(B)\big).$$

Therefore by transitivity of equality,

$$\mu^*\big(A \cap f^{\leftarrow}(B)\big) = \mu^*\big(A \cap g^{\leftarrow}(B)\big),$$

as required. ■

Remark 3.9. The following proposition will be used to extend the notion of independence of $\mathscr{L}_p(X, \mathfrak{S}, \mu)$ functions to independence of $L_p(X, \mathfrak{S}, \mu)$ classes, which will appear in Definition 6.10.

Proposition 3.10. *Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $(Y, \mathfrak{B})$ be a measurable space, and let f, g, h, $k : X \to Y$ be $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{B})$-measurable functions such that $f = h$ μ^*-almost everywhere on X and $g = k$ μ^*-almost everywhere on X. If $\mu^*(A \cap B) = \mu^*(A)\mu^*(B)$ for all $A \in f^{\leftarrow}(\mathfrak{B})$ and all $B \in g^{\leftarrow}(\mathfrak{B})$, then $\mu^*(C \cap D) = \mu^*(C)\mu^*(D)$ for all $C \in h^{\leftarrow}(\mathfrak{B})$ and all $D \in k^{\leftarrow}(\mathfrak{B})$*

Proof. Let E, $F \in \mathfrak{B}$. Since $f = h$ μ^*-almost everywhere, by Proposition 3.8,

$$\mu^*\big(f^{\leftarrow}(E) \cap k^{\leftarrow}(F)\big) = \mu^*\big(h^{\leftarrow}(E) \cap k^{\leftarrow}(F)\big)$$

and

$$\mu^*\big(f^{\leftarrow}(E)\big) = \mu^*\big(h^{\leftarrow}(E)\big).$$

Since $g = k$ μ^*-almost everywhere, again by Proposition 3.8,

$$\mu^*\big(f^{\leftarrow}(E) \cap k^{\leftarrow}(F)\big) = \mu^*\big(f^{\leftarrow}(E) \cap g^{\leftarrow}(F)\big)$$

and

$$\mu^*\big(k^{\leftarrow}(F)\big) = \mu^*\big(g^{\leftarrow}(F)\big).$$

By hypothesis,

$$\mu^*\big(f^{\leftarrow}(E) \cap g^{\leftarrow}(F)\big) = \mu^*\big(f^{\leftarrow}(E)\big)\mu^*\big(g^{\leftarrow}(F)\big),$$

and so by transitivity of equality,

$$\mu^*\big(h^{\leftarrow}(E) \cap k^{\leftarrow}(F)\big) = \mu^*\big(h^{\leftarrow}(E)\big)\mu^*\big(k^{\leftarrow}(F)\big),$$

as required. ■

Proposition 3.11. *Let $(X, \mathfrak{A}, \mu)$ be a measure space, where $\mathfrak{A}$ is a σ-algebra, and let $(Y, \mathfrak{B})$ be a measurable space. Let the function $\phi : X \to Y$ be $(\mathfrak{A}, \mathfrak{B})$-measurable. If a subset B of Y is a $(\mu \circ \phi^{\leftarrow})^*$-null set, then the subset $\phi^{\leftarrow}(B)$ of X is a μ^*-null set.*

Proof. Let B be a subset of Y which is a $(\mu \circ \phi^{\leftarrow})^*$-null set. Since $\mu \circ \phi^{\leftarrow}$ is a measure, by Proposition 3.5, it follows that the triple $(Y, \mathfrak{B}, \mu \circ \phi^{\leftarrow})$ is a measure space. So by Proposition 2.22, we can take a superset C of B with C in the σ-algebra $\mathfrak{B}$ and having the same outer measure: $(\mu \circ \phi^{\leftarrow})^*(C) = (\mu \circ \phi^{\leftarrow})^*(B) = 0$. Because the outer measure agrees with the measure on the generating semiring (σ-algebra in this case), and because ϕ is assumed measurable, meaning $\phi^{\leftarrow}(C) \in \mathfrak{A}$, the following equalities hold:

$$\begin{aligned}
(\mu \circ \phi^{\leftarrow})^*(C) &= (\mu \circ \phi^{\leftarrow})(C) && C \in \mathfrak{B}, \\
&= \mu\big(\phi^{\leftarrow}(C)\big) && \text{definition of induced measure}, \\
&= \mu^*\big(\phi^{\leftarrow}(C)\big) && \phi^{\leftarrow}(C) \in \mathfrak{A}.
\end{aligned}$$

With $B \subseteq C$, it follows that $\phi^{\leftarrow}(B) \subseteq \phi^{\leftarrow}(C)$ and since outer measure is monotone, $\mu^*\big(\phi^{\leftarrow}(B)\big) \leq \mu^*\big(\phi^{\leftarrow}(C)\big)$. The comparison $0 \leq \mu^*\big(\phi^{\leftarrow}(B)\big) \leq \mu^*\big(\phi^{\leftarrow}(C)\big) = 0$ then shows that $\phi^{\leftarrow}(B)$ is a μ^*-null subset of X, as required. ■

Proposition 3.12. *Let $(X, \mathfrak{A}, \mu)$ be a measure space, where $\mathfrak{A}$ is a σ-algebra, and let $(Y, \mathfrak{B})$ be a measurable space. Let the function $\phi : X \to Y$ be $(\mathfrak{A}, \mathfrak{B})$-measurable, and let B be a $(\mu \circ \phi^{\leftarrow})^*$-measurable subset of Y with $(\mu \circ \phi^{\leftarrow})^*(B) < \infty$. Then $\phi^{\leftarrow}(B)$ is a μ^*-measurable subset of X, and $\mu^*\big(\phi^{\leftarrow}(B)\big) = (\mu \circ \phi^{\leftarrow})^*(B)$.*

Proof. As in the proof of Proposition 3.11, and according to Proposition 2.22, we can take a superset C of B with $C \in \mathfrak{B}$ and having the same outer measure:

$$(\mu \circ \phi^{\leftarrow})^*(C) = (\mu \circ \phi^{\leftarrow})^*(B).$$

With $C \in \mathfrak{B}$, it follows that C is $(\mu \circ \phi^{\leftarrow})^*$-measurable by Proposition 2.16. Since B is a subset of C, we can write C as the disjoint union

$$C = B \sqcup (C \cap B^{\mathsf{C}}).$$

Because both C (by construction) and B (by hypothesis) are $(\mu \circ \phi^{\leftarrow})^*$-measurable, it follows that their set difference $C \cap B^{\mathsf{C}}$ is also $(\mu \circ \phi^{\leftarrow})^*$-measurable. And the outer measure $(\mu \circ \phi^{\leftarrow})^*$ is additive on the σ-algebra $(\mu \circ \phi^{\leftarrow})^*$-$\mathfrak{Meas}$, so

$$(\mu \circ \phi^{\leftarrow})^*(C) = (\mu \circ \phi^{\leftarrow})^*(B) + (\mu \circ \phi^{\leftarrow})^*(C \cap B^{\mathsf{C}}).$$

We have assumed that $(\mu \circ \phi^{\leftarrow})^*(B) < \infty$, and by construction $(\mu \circ \phi^{\leftarrow})^*(C) = (\mu \circ \phi^{\leftarrow})^*(B)$, so $(\mu \circ \phi^{\leftarrow})^*(C) < \infty$. It follows that $(\mu \circ \phi^{\leftarrow})^*(C \cap B^{\mathsf{C}}) = 0$. This would not follow were $(\mu \circ \phi^{\leftarrow})^*(B) = \infty$

Since $C \cap B^{\mathsf{C}}$ is a $(\mu \circ \phi^{\leftarrow})^*$-null set of Y, it follows by Proposition 3.11 that $\phi^{\leftarrow}(C \cap B^{\mathsf{C}})$ is a μ^*-null set of X, and therefore μ^*-$\mathfrak{Meas}$. And since $\phi^{\leftarrow}(C)$ is μ^*-$\mathfrak{Meas}$, it follows that the set difference $\phi^{\leftarrow}(B)$ of $\phi^{\leftarrow}(C)$ with $\phi^{\leftarrow}(C \cap B^{\mathsf{C}})$ is μ^*-$\mathfrak{Meas}$. We can also write $\phi^{\leftarrow}(C)$ as the disjoint union

$$\phi^{\leftarrow}(C) = \phi^{\leftarrow}(B) \sqcup \phi^{\leftarrow}(C \cap B^{\mathsf{C}}).$$

The outer measure μ^* is additive on these μ^*-$\mathfrak{Meas}$ sets, so

$$\mu^*\big(\phi^{\leftarrow}(C)\big) = \mu^*\big(\phi^{\leftarrow}(B)\big) + \mu^*\big(\phi^{\leftarrow}(C \cap B^{\complement})\big).$$

But $\phi^{\leftarrow}(C \cap B^{\complement})$ is a μ^*-null set, so

$$\mu^*\big(\phi^{\leftarrow}(C)\big) = \mu^*\big(\phi^{\leftarrow}(B)\big).$$

Also, as displayed in the proof of Proposition 3.11,

$$(\mu \circ \phi^{\leftarrow})^*(C) = \mu^*\big(\phi^{\leftarrow}(C)\big).$$

It follows by transitivity of equality that

$$\mu^*\big(\phi^{\leftarrow}(B)\big) = (\mu \circ \phi^{\leftarrow})^*(B),$$

as required. ■

Remark 3.13. One of the difficulties in working with σ-subalgebras is that the restriction of a σ-finite measure to a σ-subalgebra need not be σ-finite. This is a clear indication that our definitions of σ-finite measures and/or σ-subalgebras need more work. A definition that does not respect sub-objects must not yet be right.

Proposition 3.14. *Let $(X, \mathfrak{A}, \mu)$ be a finite measure space, where $\mathfrak{A}$ is a σ-algebra. If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, then $\mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas} \subseteq \mu^*$-$\mathfrak{Meas}$, and $\mu_{|\mathfrak{B}}^*(A) = \mu^*(A)$ for all $A \in \mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$.*

Proof. Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$, and let $A \in \mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$. The function $i : X \to X : A \mapsto A$, where $i^{\leftarrow} : \mathfrak{B} \to \mathfrak{A}$, is $(\mathfrak{A}, \mathfrak{B})$-measurable since $\mathfrak{B}$ is supposed to be a σ-subalgebra of $\mathfrak{A}$. In order to apply Proposition 3.12, we first show that $\mu_{|\mathfrak{B}} = \mu \circ i^{\leftarrow}$ on $\mathfrak{B}$: If $B \in \mathfrak{B}$, then $\mu_{|\mathfrak{B}}(B) = \mu(B) = \mu\big(i^{\leftarrow}(B)\big) = (\mu \circ i^{\leftarrow})(B)$. And so by Proposition 2.21, it follows that $\mu_{|\mathfrak{B}}^* = (\mu \circ i^{\leftarrow})^*$ on $\mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas} = (\mu \circ i^{\leftarrow})^*$-$\mathfrak{Meas}$.

Since $A \in \mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas} = (\mu \circ i^{\leftarrow})^*\text{-}\mathfrak{Meas}$, and since $(X, \mathfrak{A}, \mu)$ is finite, it follows by Proposition 3.12 that $i^{\leftarrow}(A) = A$ is a μ^*-measurable set. This shows that $\mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas} \subseteq \mu^*\text{-}\mathfrak{Meas}$. It also follows by Proposition 3.12 that $\mu^*\big(i^{\leftarrow}(A)\big) = (\mu \circ i^{\leftarrow})^*(A)$, which shows that $\mu^*(A) = \mu_{|\mathfrak{B}}^*(A)$, as required. ■

Remark 3.15. The following proposition says that two functions which are equal almost everywhere are either simultaneously measurable or not. As it should be.

Proposition 3.16. *If a function $f : X \to \mathbb{R}$ is $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, and a function g is equal to f μ^*-almost everywhere, then g is also $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable.*

Theorem 3.17. *Let $\{f_n\}$ be a sequence of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable functions. If $f : X \to \mathbb{R}$ is a function such that $f_n \to f$ μ^*-almost everywhere, then f is also $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable.*

> Underlying the mechanics of Proposition 3.16 and Theorem 3.17 is the property of completeness. This property describes a tight interplay between a measure and the collection of sets which forms its domain; let's say a measure space is *complete* if every subset of a set of measure zero is measurable. It is almost worth ferreting this out, and then restating the proposition and theorem in these terms. Vaguely, let $(X, \mathfrak{A}, \mu)$ be a complete measure space, where $\mathfrak{A}$ is at least an algebra of sets. Then if $f : X \to \mathbb{R}$ is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function, and $g : X \to \mathbb{R}$ is a function with $g = f$ μ-almost everywhere, then g is also an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function. Since the measure space $(X, \mu^*\text{-}\mathfrak{Meas}, \mu^*)$ *is* a complete measure space, the proposition would follow. But this is a type of interference that generates problems which I do not want to solve here.

Proposition 3.18. *The collection of all $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable functions forms a vector lattice.*

Definition 3.19. Let $(X, \mathfrak{A})$ be a measurable space. An $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function having finite range is called an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-**simple function**.

Proposition 3.20. *Let $(X, \mathfrak{A})$ be a measurable space, and let $f : X \to \mathbb{R}$ be an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function such that $f(x) \geq 0$ for all $x \in X$. There is a sequence $\{\phi_n\}$ of $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple functions such that $\phi_n(x) \geq 0$ and $\phi_n(x) \uparrow f(x)$ for all $x \in X$.*

Proposition 3.21. *Let $(X, \mathfrak{A})$ be a measurable space. A function $f : X \to \mathbb{R}$ is $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable if and only if there is a sequence $\{\phi_n\}$ of $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple functions such that $\lim \phi_n(x) = f(x)$ for all $x \in X$.*

Theorem 3.22. *Let $(X, \mathfrak{A}, \mu)$ be a σ-finite measure space, where $\mathfrak{A}$ is a σ-algebra. If $f : X \to \mathbb{R}$ is a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function, then there exists an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function $g : X \to \mathbb{R}$ with $f = g$ μ^*-almost everywhere.*

Proof. Roughly, we show the result holds for characteristic functions, then for simple functions, then for pointwise limits of simple functions. Assume $f(x) \geq 0$ for all $x \in X$.

Let A be a μ^*-measurable subset of X with $f = \chi_A$, so f is an $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable characteristic function. By Proposition 2.23, there is a subset $C \subseteq X$ with $A \cap C = \varnothing$ and such that $\mu^*(C) = 0$ with $A \cup C \in \mathfrak{A}$. Take $g = \chi_{A \cup C}$, so g is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable characteristic function. Then $g - f = \chi_C$, and $\mu^*(C) = 0$, which says that $f = g$ μ^*-almost everywhere.

Now let f be a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-simple function, so by definition we can let the set $\{a_1, \ldots, a_n\}$ denote the range of f. Further, let $A_i = f^{\leftarrow}(a_i)$, so each A_i is a μ^*-measurable subset of X. We can write $f = \sum_{i=1}^{n} a_i \chi_{A_i}$. Again, by Proposition 2.23, there are subsets

$C_i \subseteq X$ with $A_i \cap C_i = \varnothing$ and such that $\mu^*(C_i) = 0$ with $A_i \cup C_i \in \mathfrak{A}$. Now take $g = \sum_{i=1}^n a_i \chi_{A_i \cup C_i}$, so g is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple function. The difference $g - f = \sum_{i=1}^n a_i \chi_{C_i}$. Since $\mu^*(C_i) = 0$, this says that $g - f = 0$ μ^*-almost everywhere; equivalently, $f = g$ μ^*-almost everywhere.

Finally, let $f : X \to \mathbb{R}$ be any $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function. By Proposition 3.20, there is a sequence $\{\phi_n\}$ of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-simple functions such that $\phi_n(x) \geq 0$ and $\phi_n(x) \uparrow f(x)$ for all $x \in X$. And we have just shown that for each of these ϕ_n, there is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple function, call it y_n, such that $\phi_n = y_n$ μ^*-almost everywhere. That is, $\phi_n(x) = y_n(x)$ for all x in the complement of some set, call it A_n, with $\mu^*(A_n) = 0$. The fact that $\mu^*(A_n) = 0$ only says that $A_n \in \mu^*\text{-}\mathfrak{Meas}$. However, by Proposition 2.22, there is a superset $B_n \in \mathfrak{A}$ with $A_n \subseteq B_n$ and $\mu^*(A_n) = \mu^*(B_n)$. Then $\phi_n(x) = y_n(x)$ for all x in the complement of B_n with $\mu^*(B_n) = 0$ and $B_n \in \mathfrak{A}$. Write $B = \bigcup B_n$, so $B \in \mathfrak{A}$. It follows that $y_n(x) \uparrow f(x)$ for all x in the complement of B. Consequently, $y_n \chi_{B^c} \uparrow f\chi_{B^c}$ on all of X. This time take $g = f\chi_{B^c}$. Then $f = g$ μ^*-almost everywhere. To complete the proof, we need only show that g is $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. See that $B \in \mathfrak{A}$ implies $B^c \in \mathfrak{A}$, so χ_{B^c} is $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. Each y_n is $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, and so the product $y_n \chi_{B^c}$ is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function and consequently an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple function. Therefore by Proposition 3.21, the function g is $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, as required. ■

Definition 3.23. A $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-simple function is called a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-**step function** if it can be expressed as a linear combination of characteristic functions of μ^*-measurable sets of finite measure. For example, $\phi = \sum_{i=1}^n a_i \chi_{A_i}$, where each A_i is a μ^*-measurable set with $\mu^*(A_i) < \infty$, where we can suppose the A_i to be pairwise disjoint.

Proposition 3.24. *If $(X, \mathfrak{S}, \mu)$ is a σ-finite measure space, and $f : X \to \mathbb{R}$ is a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function with $f(x) \geq 0$ on X, then there is a sequence $\{\phi_n\}$ of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$ step functions such that $\phi_n(x) \geq 0$ and $\phi_n(x) \uparrow f(x)$ for all $x \in X$.*

Definition 3.25. Let $\phi : X \to \mathbb{R}$ be a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step function such that $\phi = \sum_{i=1}^{n} a_i \chi_{A_i}$, where the range of ϕ is the set $\{a_1, \ldots, a_n\}$, and each $A_i = \phi^{\leftarrow}(\{a_i\})$. Then the Lebesgue *integral* of ϕ, denoted by $\int_X \phi \, d\mu$, or simply $\int \phi$, is defined to be the real number

$$\int_X \phi \, d\mu = \sum_{i=1}^{n} a_i \mu^*(A_i).$$

Proposition 3.26. *Let $\phi : X \to \mathbb{R}$ be a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step function.*

1. *If $\phi \geq 0$ μ^*-almost everywhere, then $\int_X \phi \, d\mu \geq 0$.*
2. *If $\phi = 0$ μ^*-almost everywhere, then $\int_X \phi \, d\mu = 0$.*

Proposition 3.27. *Let $\{\phi_n\}$ be a sequence of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step functions. If $\phi_n \downarrow 0$ μ^*-almost everywhere, then $\int \phi_n \downarrow 0$.*

Proposition 3.28. *Let $f : X \to \mathbb{R}$. If $\{\phi_n\}$ and $\{y_n\}$ are sequences of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step functions with $\phi_n \uparrow f$ μ^*-almost everywhere, and with $y_n \uparrow f$ μ^*-almost everywhere, then $\lim \int_X \phi_n \, d\mu = \lim \int_X y_n \, d\mu$.*

Definition 3.29. Let $f : X \to \mathbb{R}$. If there is a sequence $\{\phi_n\}$ of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step functions such that $\phi_n \uparrow f$ μ^*-almost everywhere, and $\lim \int_X \phi_n \, d\mu < \infty$, then f is called an **upper function**, and the Lebesgue integral of f, denoted by $\int_X f \, d\mu$, or simply $\int f$, is defined to be the real number

$$\int_X f \, d\mu = \lim \int_X \phi_n \, d\mu.$$

We might also call f a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-upper function. By Proposition 3.28, the value of the Lebesgue integral of an upper function is independent of a chosen sequence of step functions $\phi_n \uparrow f$. By Theorem 3.17, every upper function is $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable.

Proposition 3.30. *If f and g are upper functions such that $f \geq g$ μ^*-almost everywhere, then $\int f \geq \int g$.*

Proposition 3.31. *Let $\{f_n\}$ be a sequence of upper functions. If $f_n \downarrow 0$ μ^*-almost everywhere, then $\int f_n \downarrow 0$.*

Definition 3.32. A function $f : X \to \mathbb{R}$ is **integrable** over X with respect to μ if there exist two upper functions g and h such that $f = g - h$ μ^*-almost everywhere. Then the Lebesgue integral of f over X with respect to μ, denoted by $\int_X f \, d\mu$, or simply $\int f$, is defined to be the real number

$$\int_X f \, d\mu = \int_X g \, d\mu - \int_X h \, d\mu.$$

The value of this integral is independent of the representation of f as a difference of upper functions. A function $f : X \to \mathbb{R}$ is *integrable over a μ^*-measurable subset A with respect to μ* if $f \cdot \chi_A$ is integrable over X with respect to μ, in which case $\int_A f \, d\mu$ is defined to be $\int_X f \cdot \chi_A \, d\mu$.

An extended real function $f : X \to [-\infty, +\infty]$ **defines an integrable function** over X with respect to μ if the function f "attains" $-\infty$ or $+\infty$, or is undefined, on a set of at most μ^*-measure zero, so it is possible to redefine the values of f on this set in order that the corresponding redefined function $X \to \mathbb{R}$ is integrable over X with respect to μ. We assume all such functions have been so redefined, and consequently lie in the domain of the Lebesgue integral, if that makes sense. Extended real functions are hard to avoid when it comes to dealing with signed measures, or even positive measures.

Proposition 3.33. *Let $f : X \to \mathbb{R}$ be a function integrable over X with respect to μ. If $f \geq 0$ μ^*-almost everywhere, then f is an upper function.*

Proposition 3.34. *Let $(X, \mathfrak{A}, \mu)$ be a σ-finite measure space, where $\mathfrak{A}$ is a σ-algebra. If $f : X \to \mathbb{R}$ is a function integrable over X with respect to μ, then there is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function $g : X \to \mathbb{R}$ with $f = g$ μ^*-almost everywhere.*

Proof. Let $f : X \to \mathbb{R}$ be integrable over X with respect to μ. By Theorem 3.17, the function f is $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, and by Theorem 3.22, the function f is μ^*-almost everywhere equal to an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function, as required. ■

Remark 3.35. The collection of real functions on X that can be written as the difference of upper functions forms a vector lattice, and the integral is a positive linear functional on this vector lattice. Maybe the integral here should have been defined as a positive linear functional with a continuity condition at zero, per Daniell. Of which, the Lebesgue, and so the Riemann, integral is simply a concrete example.

Proposition 3.36. *If $f : X \to \mathbb{R}$ be an integrable function, and if $g : X \to \mathbb{R}$ is a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function with $0 \leq g \leq f$ μ^*-almost everywhere, then g is an integrable function.*

Proposition 3.37. *Let $f : X \to \mathbb{R}$ be a function integrable over X with respect to μ. Then $\int |f| \, d\mu = 0$ if and only if $f = 0$ μ^*-almost everywhere.*

4 Function Spaces and Quotient Spaces

IN ORDER to deal with the L_p spaces in a manner that is actually mathematically correct, we will use three theorems from algebra; namely, the induced homomorphism theorem, the correspondence theorem, and the induced quotient homomorphism corollary. These appear respectively as Proposition 4.3, Proposition 4.5, and Proposition 4.18. These propositions are couched in terms of group theory because that is all it takes to get to the heart of the matter. Understand that each of these three theorems has a straightforward extension to modules or vector spaces.

We pay attention to the difference between a function and an equivalence class of functions for at least two reasons. The first reason is that clarity is more important here. The second reason is that we will deal with so-called *versions* of maps defined on σ-algebras which will require selecting representatives from equivalence classes of functions. These will appear in Definition 8.11.

Definition 4.1. Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $p \in (0, \infty)$. The collection of all $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable functions $f : X \to \mathbb{R}$ for which $|f|^p$ is integrable forms a semi-normed vector space, denoted by $\mathscr{L}_p(X, \mathfrak{S}, \mu)$. If $f \in \mathscr{L}_p(X, \mathfrak{S}, \mu)$, then $\int |f|^p \, d\mu = 0$ if and only if $f = 0$ μ^*-almost everywhere. A quotient space defined by identifying those functions in $\mathscr{L}_p(X, \mathfrak{S}, \mu)$ which agree μ^*-almost everywhere forms a normed vector space, denoted by $L_p(X, \mathfrak{S}, \mu)$,

where the norm will be defined in terms of the integral. That is, let $\mathscr{N}_p(X, \mathfrak{S}, \mu)$ denote the subspace of functions in $\mathscr{L}_p(X, \mathfrak{S}, \mu)$ which are equal to zero μ^*-almost everywhere, then

$$L_p(X, \mathfrak{S}, \mu) := \mathscr{L}_p(X, \mathfrak{S}, \mu)/\mathscr{N}_p(X, \mathfrak{S}, \mu).$$

For each $f \in \mathscr{L}_p(X, \mathfrak{S}, \mu)$, we will let $\overline{f}^\mu$ denote the equivalence class $f + \mathscr{N}_p(X, \mathfrak{S}, \mu)$ in $L_p(X, \mathfrak{S}, \mu)$. Should $(X, \mathfrak{S}, \mu)$ and $(X, \mathfrak{T}, \nu)$ both be measure spaces, and should the function $f : X \to \mathbb{R}$ be in both $\mathscr{L}_p(X, \mathfrak{S}, \mu)$ and $\mathscr{L}_p(X, \mathfrak{T}, \nu)$, then the equivalence class $\overline{f}^\mu$ need not be related to the equivalence class $\overline{f}^\nu$. Not even when $\mathfrak{T} \subseteq \mathfrak{S}$ and $\nu = \mu_{|\mathfrak{T}}$, as shown in Example 4.15. Should a discussion or proposition involve only a single measure μ, we might inconsistently drop the superscript and denote $\overline{f}^\mu$ by $\overline{f}$.

Since $\mathscr{N}_1(X, \mathfrak{S}, \mu)$ is in the kernel of the integral $\int$, there is a unique operator $\bar{\int}$ which makes the following diagram commute:

$$\begin{array}{ccc} \mathscr{L}_1(X, \mathfrak{S}, \mu) & \xrightarrow{\int} & \mathbb{R} \\ {\scriptstyle \text{quotient}}\downarrow & \nearrow_{\bar{\int}} & \\ L_1(X, \mathfrak{S}, \mu) & & \end{array}$$

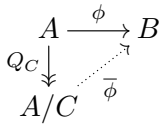
It means that for a class $\overline{f}^\mu \in L_p(X, \mathfrak{S}, \mu)$, the value of $\bar{\int}_X \overline{f}^\mu \, d\mu$ is defined to be equal to the value of $\int_X f \, d\mu$. Verify that the map $\bar{\int}$ is constant on the cosets of $L_p(X, \mathfrak{S}, \mu)$. The integral symbol '$\bar{\int}$' with the overbar was chosen to somewhat distinguish itself from the plain integral symbol '$\int$'. We have a very large number of symbols in mathematics. Do not use the exact same symbol for everything. As usual, define the L_p norm by

$$\left\| \overline{f}^\mu \right\|_p := \left(\int |f|^p \, d\mu \right)^{1/p}.$$

The quotient space $L_p(X, \mathfrak{S}, \mu)$ need not form an algebra with respect to pointwise multiplication. Still, for classes $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ in $L_p(X, \mathfrak{S}, \mu)$, if the function $|fg|^p$ is integrable, then we can define the product $\overline{f}^{\mu}\overline{g}^{\mu}$ of classes to be the class $\overline{fg}^{\mu}$. In this case, verify that if $h \in \overline{f}^{\mu}$, and $k \in \overline{g}^{\mu}$, then $\overline{hk}^{\mu} = \overline{fg}^{\mu}$.

Remark 4.2. The extension of the following proposition to modules, or vector spaces, underlies the commutative diagram in Definition 4.1.

Proposition 4.3. *If A and B are groups with $C \lhd A$, and ϕ is a homomorphism $A \to B$ with C contained in the $\operatorname{Ker}\phi$, there is a unique induced homomorphism $\overline{\phi} : A/C \to B$ mapping the coset $a + C$ to $\phi(a)$, so that the following diagram commutes.*



The $\operatorname{Ran}\overline{\phi} = \operatorname{Ran}\phi$ and the $\operatorname{Ker}\overline{\phi} = (\operatorname{Ker}\phi)/C$. The induced homomorphism is an isomorphism exactly when both ϕ is epic and $C = \operatorname{Ker}\phi$.

Remark 4.4. The $L_p(X, \mathfrak{S}, \mu)$ spaces can be comparable; for example, let $(X, \mathfrak{S}, \mu)$ be a finite measure space, then the quotient space $L_2(X, \mathfrak{S}, \mu)$ is a vector subspace of the quotient space $L_1(X, \mathfrak{S}, \mu)$. This is the heart of the correspondence theorem, stated in Proposition 4.5, which extends to vector spaces, and it would be necessary that $\mathscr{N}_2(X, \mathfrak{S}, \mu) = \mathscr{N}_1(X, \mathfrak{S}, \mu)$ in order for the proposition to apply. The result then that $L_2(X, \mathfrak{S}, \mu)$ is a vector subspace of $L_1(X, \mathfrak{S}, \mu)$ follows from the fact that for a finite measure space, $\mathscr{L}_2(X, \mathfrak{S}, \mu) \subseteq \mathscr{L}_1(X, \mathfrak{S}, \mu)$, which in turn is an application of Hölder's inequality.

Proposition 4.5. *If $f : G \twoheadrightarrow H$ is an epimorphism of groups, then the assigment $K \mapsto f(K)$ defines a one-to-one correspondence between the set of all subgroups K of G which contain the $\operatorname{Ker} f$ and the set of all subgroups of H.*

$$\begin{array}{ccccc} \operatorname{Ker} f & \leq & K & \leq & G \\ & & \downarrow & & \downarrow \\ & & f(K) & \leq & H \end{array}$$

Consequently, if N is a normal subgroup of a group G, then every subgroup of G/N is of the form K/N, where K is a subgroup of G that contains N.

$$\begin{array}{ccccc} N & \leq & K & \leq & G \\ \downarrow & & \downarrow & & \downarrow \\ N/N & \leq & K/N & \leq & G/N \end{array}$$

Proposition 4.6. *If $(X, \mathfrak{S}, \mu)$ is a measure space, then $\mathscr{N}_2(X, \mathfrak{S}, \mu) = \mathscr{N}_1(X, \mathfrak{S}, \mu)$.*

Proof. If $f \in \mathscr{L}_2(X, \mathfrak{S}, \mu)$ and $f = 0$ μ^*-almost everywhere, then $f \in \mathscr{L}_1(X, \mathfrak{S}, \mu)$, and if $f \in \mathscr{L}_1(X, \mathfrak{S}, \mu)$ and $f = 0$ μ^*-almost everywhere, then $f \in \mathscr{L}_2(X, \mathfrak{S}, \mu)$. ■

Proposition 4.7. *If $(X, \mathfrak{S}, \mu)$ is finite, then $L_2(X, \mathfrak{S}, \mu)$ is a vector subspace of the quotient space $L_1(X, \mathfrak{S}, \mu)$.*

$$\begin{array}{ccccc} \mathscr{N}_1(X, \mathfrak{A}, \mu) & \leq & \mathscr{L}_2(X, \mathfrak{A}, \mu) & \leq & \mathscr{L}_1(X, \mathfrak{A}, \mu) \\ \downarrow & & \downarrow & & \downarrow \\ \textit{trivial} & \leq & L_2(X, \mathfrak{A}, \mu) & \leq & L_1(X, \mathfrak{A}, \mu) \end{array}$$

Proposition 4.8. *Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let f be an integrable function such that $f > 0$ μ^*-almost everywhere. If A is a μ^*-measurable subset of X, and $\int_A f \, d\mu = 0$, then $\mu^*(A) = 0$.*

Proposition 4.9. *Let $(X, \mathfrak{S}, \mu)$ be a measure space. If the class $\overline{f}^{\mu} \in L_1(X, \mathfrak{S}, \mu)$, and if $\int_A f\, d\mu = 0$ for all $A \in \mathfrak{S}$, then $\overline{f}^{\mu} = \overline{0}^{\mu}$. Alternately, if $\overline{f}^{\mu}$, $\overline{g}^{\mu} \in L_1(X, \mathfrak{S}, \mu)$, and if $\int_A f\, d\mu = \int_A g\, d\mu$ for all $A \in \mathfrak{S}$, then $\overline{f}^{\mu} = \overline{g}^{\mu}$.*

Compare the following Proposition 4.10 (where we assume integrability) with Proposition 3.24 (where we do not assume integrability, but we do assume σ-finiteness):

> Proposition 3.24. If $(X, \mathfrak{S}, \mu)$ is a σ-finite measure space, and $f : X \to \mathbb{R}$ is a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function with $f(x) \geq 0$ on X, then there is a sequence $\{\phi_n\}$ of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$ step functions such that $\phi_n(x) \geq 0$ and $\phi_n(x) \uparrow f(x)$ for all $x \in X$.

Proposition 4.10. *For every $p \in [1, \infty)$, the equivalence classes of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step functions form a vector sublattice of the Banach lattice $L_p(X, \mathfrak{S}, \mu)$, which is norm dense.*

Proposition 4.11. *If E is a Banach lattice and F is a normed vector lattice, then every positive operator $E \to F$ is continuous.*

Remark 4.12. The integral $\bar{\int}$ on $L_1(X, \mathfrak{S}, \mu)$ is an example of a positive operator on a Banach lattice, and so the integral on $L_1(X, \mathfrak{S}, \mu)$ is necessarily continuous. But with respect to which topologies?

The following proposition describes what is sometimes called a *change of variable*, and perhaps should have been couched as:

$$(X, \mathfrak{A}, \mu) \xrightarrow[T]{} (Y, \mathfrak{B}, \mu \circ T^{\leftarrow}) \xrightarrow{f} (\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}), \qquad \overset{f\circ T}{\frown}$$
$$T^{\leftarrow}(B) \qquad\qquad B$$

with

$$\int_{T^{\leftarrow}(B)} (f \circ T)\, d\mu = \int_B f\, d(\mu \circ T^{\leftarrow}).$$

Proposition 4.13 (Change of Variable). *Let $(X, \mathfrak{A}, \mu)$ be a measure space, where $\mathfrak{A}$ is a σ-algebra, and let $(Y, \mathfrak{B})$ be a measurable space. Also, let the function $T : X \to Y$ be $(\mathfrak{A}, \mathfrak{B})$-measurable, and momentarily let $\nu = \mu \circ T^{\leftarrow}$. If $\overline{f}^{\nu} \in L_1(Y, \mathfrak{B}, \nu)$, then $\overline{f \circ T}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$ and*

$$\int_X (f \circ T)\, d\mu = \int_Y f\, d\nu.$$

Proof. This proof relies upon the fact that $\chi_B \circ T = \chi_{T^{\leftarrow}(B)}$. We will show that the displayed equality holds in case $f = \chi_B$, where we let the subset B of Y be a $(\mu \circ T^{\leftarrow})^*$-measurable set with $(\mu \circ T^{\leftarrow})^*(B) < \infty$, and note that the outer measure $(\mu \circ T^{\leftarrow})^*(B)$ should be finite in order that χ_B be integrable with respect to the induced measure $\mu \circ T^{\leftarrow}$.

By Proposition 3.12, the set $T^{\leftarrow}(B)$ is μ^*-measurable with

$$\mu^*\big(T^{\leftarrow}(B)\big) = (\mu \circ T^{\leftarrow})^*(B),$$

and so

$$\begin{aligned}
\int_X \chi_B \circ T\, d\mu &= \int_X \chi_{T^{\leftarrow}(B)}\, d\mu && \chi_B \circ T = \chi_{T^{\leftarrow}(B)} \\
&= \mu^*\big(T^{\leftarrow}(B)\big) && \text{by Definition 3.25} \\
&= (\mu \circ T^{\leftarrow})^*(B) && \text{Proposition 3.12} \\
&= \int_Y \chi_B\, d(\mu \circ T^{\leftarrow}) && \text{by Definition 3.25,}
\end{aligned}$$

as claimed. ∎

Proposition 4.14. *Let $(X, \mathfrak{A}, \mu)$ be a finite measure space, where $\mathfrak{A}$ is a σ-algebra, and $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$. If $\overline{f}^{\mu_{|\mathfrak{B}}} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$,*

then $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$ *and*

$$\int_X f \, d\mu_{|\mathfrak{B}} = \int_X f \, d\mu.$$

Consequently, the operator

$$\mathfrak{I} : L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \to L_1(X, \mathfrak{A}, \mu)$$

defined by $\overline{f}^{\mu_{|\mathfrak{B}}} \mapsto \overline{f}^{\mu}$ *has norm 1. Furthermore, the operator* $\mathfrak{I}$ *is injective.*

Proof. Let $\overline{f}^{\mu_{|\mathfrak{B}}} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. The idea is to apply a change of variable. Define the function $i : X \to X : x \mapsto x$, where $i^{\leftarrow} : \mathfrak{B} \to \mathfrak{A} : B \mapsto B$. The function i is $(\mathfrak{A}, \mathfrak{B})$-measurable because $\mathfrak{B}$ is supposed to be a σ-subalgebra of $\mathfrak{A}$.

We now show that $\int_X f \circ i \, d\mu = \int_X f \, d(\mu \circ i^{\leftarrow})$ by applying Proposition 4.13 to the measure space $(X, \mathfrak{A}, \mu)$ and the measurable space $(X, \mathfrak{B})$ and the function i. Since i is $(\mathfrak{A}, \mathfrak{B})$-measurable, and $\overline{f}^{\mu_{|\mathfrak{B}}} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, it follows by Proposition 4.13 that $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$ and

$$\int_X (f \circ i) \, d\mu = \int_X f \, d(\mu \circ i^{\leftarrow}).$$

We next show that $\int_X f \, d(\mu \circ i^{\leftarrow}) = \int_X f \, d\mu_{|\mathfrak{B}}$ by applying Proposition 3.14. The assumption that the measure space $(X, \mathfrak{A}, \mu)$ be finite is essential here. The measure $\mu \circ i^{\leftarrow}$ is equal to μ on $\mathfrak{B}$; that is, if $B \in \mathfrak{B}$, then $\mu \circ i^{\leftarrow}(B) = \mu\big(i^{\leftarrow}(B)\big) = \mu(B)$. This means that $\mu \circ i^{\leftarrow}$ and μ generate the same outer measure on the σ-algebra $\mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$. That is, $(\mu \circ i^{\leftarrow})^* = \mu^*$ on $\mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$. Because the measure space $(X, \mathfrak{A}, \mu)$ is finite, it follows by Proposition 3.14 that $\mu^* = \mu_{|\mathfrak{B}}^*$ on $\mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$. Since $(\mu \circ i^{\leftarrow})^* = \mu_{|\mathfrak{B}}^*$ on the σ-algebra $\mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$, it now follows by the definition of the integral that

$$\int_X f \, d(\mu \circ i^{\leftarrow}) = \int_X f \, d\mu_{|\mathfrak{B}}.$$

We finally see that

$$\int_X f\,d\mu = \int_X f \circ i\,d\mu,$$

and this follows from the fact that $f = f \circ i$ on X.

Therefore, by transitivity of equality,

$$\int_X f\,d\mu_{|\mathfrak{B}} = \int_X f\,d\mu \tag{4.1}$$

as required; in other words, the operator $\mathfrak{I}$ preserves the integral.

To see that $\|\mathfrak{I}\| = 1$, witness:

$$\begin{aligned}\|\mathfrak{I}\| &= \sup\big\{\|\,\mathfrak{I}\,\overline{f}^{\mu_{|\mathfrak{B}}}\| : \|\overline{f}^{\mu_{|\mathfrak{B}}}\| = 1\big\}\\ &= \sup\big\{\|\overline{f}^{\mu}\| : \|\overline{f}^{\mu_{|\mathfrak{B}}}\| = 1\big\}\\ &= \sup\{\textstyle\int |f|\ d\mu : \int |f|\ d\mu_{|\mathfrak{B}} = 1\big\}.\end{aligned}$$

And we have just barely shown in equation (4.1) that the operator $\mathfrak{I}$ preserves the integral, so $\int |f|\ d\mu = \int |f|\ d\mu_{|\mathfrak{B}}$ for all f in $\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. Since an algebra is nonempty, there is at least one f in $\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ such that $\int |f|\ d\mu_{|\mathfrak{B}} = 1$, it follows that $\|\mathfrak{I}\| = 1$.

To show that $\mathfrak{I}$ is injective, it is sufficient to show that the kernel $\mathfrak{I}^{\leftarrow}\big(\overline{0}^{\mu}\big)$ of $\mathfrak{I}$ is trivial, which means showing that $\mathfrak{I}^{\leftarrow}\big(\overline{0}^{\mu}\big) = \{\overline{0}^{\mu_{|\mathfrak{B}}}\}$ in $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.

To show the inclusion $\{\overline{0}^{\mu_{|\mathfrak{B}}}\} \subseteq \mathfrak{I}^{\leftarrow}\big(\overline{0}^{\mu}\big)$ means showing that $\mathfrak{I}(\overline{0}^{\mu_{|\mathfrak{B}}}) = \overline{0}^{\mu}$. But this holds by the very definition of $\mathfrak{I}$.

To show the inclusion $\mathfrak{I}^{\leftarrow}\big(\overline{0}^{\mu}\big) \subseteq \{\overline{0}^{\mu_{|\mathfrak{B}}}\}$, let $\overline{f}^{\mu_{|\mathfrak{B}}}$ be any element of $\mathfrak{I}^{\leftarrow}\big(\overline{0}^{\mu}\big)$, and we want to see that this implies $\overline{f}^{\mu_{|\mathfrak{B}}} = \overline{0}^{\mu_{|\mathfrak{B}}}$. To assume that $\overline{f}^{\mu_{|\mathfrak{B}}}$ is in $\mathfrak{I}^{\leftarrow}\big(\overline{0}^{\mu}\big)$ means to assume that $\mathfrak{I}\,\overline{f}^{\mu_{|\mathfrak{B}}} = \overline{0}^{\mu}$, or equivalently that $\overline{f}^{\mu} = \overline{0}^{\mu}$. In turn, this means to assume that $f \in \mathscr{N}_1(X, \mathfrak{A}, \mu)$, so that $\int |f|\ d\mu = 0$. We have just shown

in equation (4.1) that the operator $\mathfrak{I}$ preserves the integral, so $\int |f|\, d\mu = 0$ implies that $\int |f|\, d\mu_{|\mathfrak{B}} = 0$. This says $f \in \mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, which also says $\overline{f}^{\mu_{|\mathfrak{B}}} = \overline{0}^{\mu_{|\mathfrak{B}}}$, as required. ■

Example 4.15. Let us show by way of example that $\mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ need not equal $\mathscr{N}_1(X, \mathfrak{A}, \mu)$. Let $\mathfrak{A} = \{\varnothing, A, A^{\mathsf{c}}, X\}$ with Dirac measure δ_x concentrated at a point $x \in A$, with A non-empty, and with A not equal to X. Let $\mathfrak{B} = \{\varnothing, X\}$. Verify that $\delta_x^*(A) = \delta_x(A) = 1$, and $\delta_x^*(A^{\mathsf{c}}) = \delta_x(A^{\mathsf{c}}) = 0$. Then $\chi_{A^{\mathsf{c}}} = 0$ δ_x^*-almost everywhere. That is, the set of points in X where $\chi_{A^{\mathsf{c}}}$ is not equal to zero, namely the set A^{c}, is a set of δ_x^*-measure zero. Since $\chi_{A^{\mathsf{c}}}$ is $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, and $\chi_{A^{\mathsf{c}}} = 0$ δ_x^*-almost everywhere, it follows that $\chi_{A^{\mathsf{c}}} \in \mathscr{N}_1(X, \mathfrak{A}, \delta_x)$. But ${\chi_{A^{\mathsf{c}}}}^{\leftarrow}(\{1\}) = A^{\mathsf{c}}$, and the set A^{c} is not in $\mathfrak{B}$, so the function $\chi_{A^{\mathsf{c}}}$ is not $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. Then $\chi_{A^{\mathsf{c}}}$ is not in $\mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ since every function in $\mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ is required, by definition, to be $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. Conclude that the set $\mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ need not equal the set $\mathscr{N}_1(X, \mathfrak{A}, \mu)$.

Discussion 4.16. What does Proposition 4.14 not say? It does not say that the quotient space $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ is a subspace of the quotient space $L_1(X, \mathfrak{A}, \mu)$. This is because the quotient space $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ is not in general a subspace of the quotient space $L_1(X, \mathfrak{A}, \mu)$, and this is because the kernel $\mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ need not equal the kernel $\mathscr{N}_1(X, \mathfrak{A}, \mu)$, as pointed out in Example 4.15; all possible subspaces of quotient spaces are characterized by the correspondence in Proposition 4.5.

Still, the *function* space $\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ *is* a subspace of the function space $\mathscr{L}_1(X, \mathfrak{A}, \mu)$, and the inclusion $\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \subseteq \mathscr{L}_1(X, \mathfrak{A}, \mu)$ does preserve the semi-norm, so there is an inclusion of subspaces: $\mathscr{N}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \subseteq \mathscr{N}_1(X, \mathfrak{A}, \mu)$. This implies that there is a unique map

making the following diagram commute:

$$\begin{array}{ccc} L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) & \cdots\cdots\cdots\cdots\rightarrow & L_1(X, \mathfrak{A}, \mu) \\ \text{quotient}\uparrow\!\!\uparrow & & \uparrow\!\!\uparrow\text{quotient} \\ \mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) & \xrightarrow{\text{inclusion}} & \mathscr{L}_1(X, \mathfrak{A}, \mu) \end{array}$$

The distinction between a subset and subspace could be a little less blurred. Rather than the notation

$$\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \subseteq \mathscr{L}_1(X, \mathfrak{A}, \mu)$$

something better might be

$$\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \leq \mathscr{L}_1(X, \mathfrak{A}, \mu)$$

suggesting some kind of structure-preserving thing.

Remark 4.17. The extension of the following proposition to modules or vector spaces underlies some of the mechanics of the map $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \to L_1(X, \mathfrak{A}, \mu)$.

Proposition 4.18. *Whenever A and B are groups with $C \lhd A$ and $D \lhd B$, and ϕ is a homomorphism $A \to B$ such that its restriction $\phi|_C$ is a homomorphism $C \to D$, there is a unique induced homomorphism $\widehat{\phi} : A/C \to B/D$ mapping the coset $a + C$ to the coset $\phi(a) + D$.*

$$\begin{array}{ccc} A & \xrightarrow{\phi} & B \\ \triangledown & & \triangledown \\ C & \xrightarrow{\phi|_C} & D \end{array} \quad \Longrightarrow \quad \begin{array}{ccc} A/C & \xrightarrow{\widehat{\phi}} & B/D \\ \cup & & \cup \\ a + C & \mapsto & \phi(a) + D \end{array}$$

The induced homomorphism $\widehat{\phi}$ is an isomorphism exactly when both $(\operatorname{Ran}\phi) \vee D = B$ and $\phi^{\leftarrow}(D) \leq C$.

Remark 4.19. You may be wondering if there is a more direct way of proving Proposition 4.14, possibly one which more directly uses the mechanics of Proposition 4.18, but which avoids the change of variable. Propositions 4.13 and 4.14 are really just elaborations of Propositions 3.12 and 3.14, and it will be interesting to see how you would avoid these latter two.

Remark 4.20. The following proposition is a partial converse to Proposition 4.14.

Proposition 4.21. *Let $(X, \mathfrak{A}, \mu)$ be a finite measure space, where $\mathfrak{A}$ is a σ-algebra, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. If $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, and if the function f is $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, then $\overline{f}^{\mu_{|\mathfrak{B}}} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.*

Proof. Let $f \in \mathscr{L}_1(X, \mathfrak{A}, \mu)$, and let f be $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. We can suppose that $f \geq 0$. By Proposition 3.20, there is a sequence $\{\phi_n\}$ of $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-simple functions such that $\phi_n(x) \geq 0$ and $\phi_n(x) \uparrow f(x)$ for all $x \in X$. In a finite measure space, every $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-simple function is a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-step function. But every $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-step function is a $(\mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step function since, by Proposition 2.16, every element of $\mathfrak{B}$ is a $\mu_{|\mathfrak{B}}^*$-measurable subset. This says $\{\phi_n\}$ is a sequence of $(\mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step functions with $\phi_n \uparrow f$ for all $x \in X$. In order to complete the proof that f is integrable with respect to $\mu_{|\mathfrak{B}}$, it is sufficient to show that $\lim \int_X \phi_n \, d\mu_{|\mathfrak{B}} < \infty$.

By Proposition 4.14, it follows that

$$\int_X \phi_n \, d\mu = \int_X \phi_n \, d\mu_{|\mathfrak{B}}$$

for each n. And each $(\mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step function is also a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step function, so the sequence $\{\phi_n\}$ is comprised of $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-step functions with $\phi_n \uparrow f$ for all $x \in X$. Since f is

a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-upper function, it follows that $\int f\,d\mu < \infty$, and it follows by the uniqueness of the integral that $\int_X f\,d\mu = \lim \int_X \phi_n\,d\mu$. Then since $\int_X \phi_n\,d\mu = \int_X \phi_n\,d\mu_{|\mathfrak{B}}$, it follows that $\lim \int_X \phi_n\,d\mu = \lim \int_X \phi_n\,d\mu_{|\mathfrak{B}} < \infty$, completing the proof. ■

Remark 4.22. Let $(X, \mathfrak{A}, \mu)$ be a measure space, where $\mathfrak{A}$ is a σ-algebra, and let $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. The representative function f is necessarily $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. By Proposition 2.16, we know that $\mathfrak{A}$ is a σ-subalgebra of $\mu^*\text{-}\mathfrak{Meas}$, implying that any $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable is also a $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function. The converse need not hold. That is, the function f need *not* also be $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. Regardless, each class in $L_1(X, \mathfrak{A}, \mu)$ *does* have an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative, as claimed in the following proposition, which is just a restatement of Proposition 3.22 in terms of the quotient space $L_1(X, \mathfrak{A}, \mu)$.

Proposition 4.23. *Let $(X, \mathfrak{A}, \mu)$ be a σ-finite measure space, where $\mathfrak{A}$ is a σ-algebra. If $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, then there is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function $g \in \mathscr{L}_1(X, \mathfrak{A}, \mu)$ such that $\overline{g}^{\mu} = \overline{f}^{\mu}$. That is, each class of $L_1(X, \mathfrak{A}, \mu)$ has an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative.*

Definition 4.24. Let ν and μ be signed measures on a σ-algebra $\mathfrak{A}$. The signed measure ν is **absolutely continuous** with respect to μ on $\mathfrak{A}$ if $|\mu|(A) = 0$ implies $\nu(A) = 0$ whenever $A \in \mathfrak{A}$. Denote this by $\nu \ll \mu$.

Proposition 4.25. *Let $(X, \mathfrak{A})$ be a measurable space, and let μ and ν be measures on $\mathfrak{A}$ with $\nu \ll \mu$.*

1. *If $A \subseteq X$ and $\mu^*(A) = 0$, then $\nu^*(A) = 0$.*
2. *If μ is σ-finite, then $\mu^*\text{-}\mathfrak{Meas} \subseteq \nu^*\text{-}\mathfrak{Meas}$.*

Definition 4.26. Let $(X, \mathfrak{A}, \mu)$ be a measure space, where $\mathfrak{A}$ is a σ-algebra. For every class $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$ there is a finite signed measure $\nu : \mathfrak{A} \to \mathbb{R}$ on the measurable space $(X, \mathfrak{A})$ defined by

$$\nu(A) = \int_A f \, d\mu \quad \text{for all } A \in \mathfrak{A}.$$

This finite signed measure ν is absolutely continuous with respect to μ on $\mathfrak{A}$. Call ν the **indefinite integral** of $\overline{f}^{\mu}$. The idea that an integrable function defines a measure that is absolutely continuous with respect to some measure has the following form of converse.

Theorem 4.27 (Radon-Nikodym). *Let $(X, \mathfrak{A})$ be a measurable space. If ν is a finite signed measure on $\mathfrak{A}$ that is absolutely continuous with respect to a σ-finite measure μ on $\mathfrak{A}$, then there exists a unique class $\overline{f}^{\mu}$ in $L_1(X, \mathfrak{A}, \mu)$ such that*

$$\nu(A) = \int_A f \, d\mu \quad \textit{for all } A \in \mathfrak{A}.$$

Furthermore, we may suppose the representative function f to be $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable.

Definition 4.28. The unique class $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$ in the previous theorem is called the **Radon-Nikodym derivative** of ν with respect to μ, and may be denoted $\overline{d\nu/d\mu}^{\mu}$. The $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative function f may be denoted $d\nu/d\mu$, and will be referred to as the **density function** of ν with respect to μ whenever having to say "a Radon-Nikodym derivative representative" becomes tiresome, which happens pretty quickly.

Sometimes, as in the case of Proposition 4.29, the notation $d\nu = f \, d\mu$ even makes sense.

Proposition 4.29. *Let $(X, \mathfrak{A})$ be a measurable space, and let ν be a finite measure on $\mathfrak{A}$ and μ be σ-finite measure on $\mathfrak{A}$ with $\nu \ll \mu$. Also let $\overline{f}^{\mu} = \overline{d\nu/d\mu}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$.*

1. *Let $A = \{x \in X : f(x) > 0\}$. If $B \in \nu^*\text{-}\mathfrak{Meas}$, then $A \cap B \in \mu^*\text{-}\mathfrak{Meas}$.*
2. *If the class $\overline{g}^{\nu} \in L_1(X, \mathfrak{A}, \nu)$, then the class $\overline{gf}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, and*
$$\int_X g \, d\nu = \int_X gf \, d\mu,$$
in which case "$d\nu = f \, d\mu$."

Proposition 4.30. *Let $(X, \mathfrak{A})$ be a measurable space, and let ν and μ be finite measures on $\mathfrak{A}$, and let λ be a σ-finite measure on $\mathfrak{A}$. If $\nu \ll \mu$ and $\mu \ll \lambda$, then $\nu \ll \lambda$ and*
$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu}\frac{d\mu}{d\lambda}.$$

Definition 4.31. Let $(Y, \mathfrak{B})$ be a measurable space, and let $f : X \to Y$ be a function, where $f^{\leftarrow} : \mathfrak{B} \to \mathfrak{P}(X)$. Then $f^{\leftarrow}(\mathfrak{B}) := \{f^{\leftarrow}(B) : B \in \mathfrak{B}\}$ is a σ-subalgebra of the power set $\mathfrak{P}(X)$; it's the *σ-algebra induced by f*. The notation $\sigma(f)$ is also used.

The function $f : X \to Y$ is always $\big(f^{\leftarrow}(\mathfrak{B}), \mathfrak{B}\big)$-measurable by the very definition of what it means for a function to be measurable. Should a σ-subalgebra $\mathfrak{A}$ of $\mathfrak{P}(X)$ be specified, so that $(X, \mathfrak{A})$ is a measurable space, and should $f : X \to Y$ be $(\mathfrak{A}, \mathfrak{B})$-measurable, then $f^{\leftarrow}(\mathfrak{B})$ is a σ-subalgebra of $\mathfrak{A}$. To indicate that $f : X \to Y$ is $(\mathfrak{A}, \mathfrak{B})$-measurable, we might use the notation $f : (X, \mathfrak{A}) \to (Y, \mathfrak{B})$.

Notation 4.32. When it comes to commutative diagrams, you have to decide what it means to commute. In the following, there will be

diagrams that commute up to sets of measure zero. For example, in the following diagram

$$\begin{array}{ccc} \big(X, \mathfrak{A}\big) & \xrightarrow{\;g\;} & (\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}) \\ & \searrow^{f} & \uparrow{\scriptstyle h} \\ & & (Y, \mathfrak{B}) \end{array}$$

if f and h are functions, and if g is a representative of an $L_1(X, \mathfrak{A}, \mu)$ class, then what does it mean to say that the diagram commutes? If we mean that it commutes up to subsets of $\mathfrak{A}$ having μ^*-measure zero, or equivalently that $\overline{g}^{\mu} = \overline{h \circ f}^{\mu}$, then we will qualify the diagram, saying just that:

$$\begin{array}{ccc} \big(X, \mathfrak{A}\big) & \xrightarrow{\;g\;} & (\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}) \\ & \searrow^{f} & \uparrow{\scriptstyle h} \\ & & (Y, \mathfrak{B}) \end{array} \qquad \text{meaning} \quad \overline{g}^{\mu} = \overline{h \circ f}^{\mu}.$$

Remark 4.33. Although the following proposition looks like the Doob-Dynkin lemma, it is not simply about measurability. It's also about a form of converse to Proposition 4.13.

The double-headed surjective arrow '$\twoheadrightarrow$' indicates that a function is surjective.

Proposition 4.34. *Let $(X, \mathfrak{A}, \mu)$ be a finite measure space, where $\mathfrak{A}$ is a σ-algebra, and let $(Y, \mathfrak{B})$ be a measurable space. Also let $f : X \to Y$ be an $(\mathfrak{A}, \mathfrak{B})$-measurable function such that $f^{\leftarrow} : \mathfrak{B} \twoheadrightarrow \mathfrak{A}$ is surjective. Let $\overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, then there is a unique class $\overline{h}^{\mu} \in L_1(Y, \mathfrak{B}, \mu \circ f^{\leftarrow})$ with $\overline{g}^{\mu} = \overline{h \circ f}^{\mu}$ and with h being a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function:*

$$\begin{array}{ccc} \big(X, \mathfrak{A}\big) & \xrightarrow{\;g\;} & (\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}) \\ & \searrow^{f} & \vdots{\scriptstyle h} \\ & & (Y, \mathfrak{B}) \end{array} \qquad \textit{meaning} \quad \overline{g}^{\mu} = \overline{h \circ f}^{\mu}$$

Proof. The idea is to use the Radon-Nikodym Theorem to get the class $\overline{h}^{\mu\circ f^{\leftarrow}} \in L_1(Y, \mathfrak{B}, \mu \circ f^{\leftarrow})$ with $\overline{g}^{\mu} = \overline{h \circ f}^{\mu}$.

Let ϕ denote the indefinite integral of g with respect to μ. That is,

$$\phi(A) = \int_A g \, d\mu \quad \text{for all } A \in \mathfrak{A}.$$

The set function $\phi \circ f^{\leftarrow} : \mathfrak{B} \to [0, \infty]$ defines a finite measure that is σ-additive and absolutely continuous with respect to the finite (and therefore σ-finite) measure $\mu \circ f^{\leftarrow}$ on $(Y, \mathfrak{B})$. By the Radon-Nikodym theorem (4.27), there is a unique class $\overline{h}^{\mu\circ f^{\leftarrow}} \in L_1(Y, \mathfrak{B}, \mu \circ f^{\leftarrow})$ such that

$$(\phi \circ f^{\leftarrow})(B) = \int_B h \, d(\mu \circ f^{\leftarrow}) \quad \text{for all } B \in \mathfrak{B},$$

and with h a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function. On the one hand, by definition,

$$\big(\phi \circ f^{\leftarrow}(B)\big) = \phi\big(f^{\leftarrow}(B)\big) = \int_{f^{\leftarrow}(B)} g \, d\mu.$$

On the other hand, by Proposition 4.13,

$$\int_B h \, d(\mu \circ f^{\leftarrow}) = \int_{f^{\leftarrow}(B)} h \circ f \, d\mu.$$

By transitivity of equality, it follows that

$$\int_{f^{\leftarrow}(B)} g \, d\mu = \int_{f^{\leftarrow}(B)} h \circ f \, d\mu \quad \text{for all } B \in \mathfrak{B}.$$

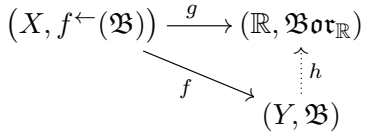
The assumption that $f^{\leftarrow}$ is surjective means that $f^{\leftarrow}(\mathfrak{B}) = \mathfrak{A}$, and so

$$\int_A g \, d\mu = \int_A h \circ f \, d\mu \quad \text{for all } A \in \mathfrak{A}.$$

By Proposition 4.9, this implies that $\overline{g}^{\mu} = \overline{h \circ f}^{\mu}$, completing the proof. ∎

For comparison with Proposition 4.34, the so-called Doob-Dynkin lemma.

Proposition 4.35 (Doob-Dynkin). *Let the space $(Y, \mathfrak{B})$ be measurable, and let $g : X \to \mathbb{R}$, and $f : X \to Y$. The function g is $(f^{\leftarrow}(\mathfrak{B}), \mathfrak{Bor}_{\mathbb{R}})$-measurable if and only if there is a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function $h : Y \to \mathbb{R}$ such that $g = h \circ f$:*



5 Product Measures and Product Integration

THROUGHOUT this section, let $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ be measure spaces, so $\mathfrak{S}$ and $\mathfrak{T}$ are at least semirings. The powerset of any set A is still denoted $\mathfrak{P}(A)$. And μ^*-$\mathfrak{Meas}$ still denotes the σ-algebra of μ^*-measurable subsets of X. As stated in Definition 2.9, we may let μ^* also denote a restriction $\mu^*_{|}$.

5.1 Product Measures

Definition 5.1. If $\mathfrak{S}$ and $\mathfrak{T}$ are semirings of subsets of respective sets X and Y, then the collection of Cartesian products

$$\{A \times B \subseteq X \times Y : A \in \mathfrak{S}, B \in \mathfrak{T}\}$$

forms a semiring of subsets of $X \times Y$, generated in the first component by $\mathfrak{S}$ and in the second component by $\mathfrak{T}$, and will consequently be denoted $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$. It may be called the **product semiring** of $\mathfrak{S}$ and $\mathfrak{T}$, however, verify that the semiring $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$ is a subset of $\mathfrak{P}(X \times Y)$, and it is not the same as the Cartesian product $\mathfrak{S} \times \mathfrak{T}$, which is a subset of $\mathfrak{P}(X) \times \mathfrak{P}(Y)$. Semirings are non-empty; they have at least one element.

The idea of a product semiring generalizes in a clear way to the product of more than two factors if we make an identification. The

Cartesian product is not associative, so we treat $(A \times B) \times C$ and $A \times (B \times C)$ as equivalent, and may (co)universally write $A \times B \times C$.

Proposition 5.2. *The set function $\mu \times \nu$ defined by*

$$\mu \times \nu : \mathfrak{Semi}(\mathfrak{S}, \mathfrak{T}) \to [0, \infty] : A \times B \mapsto \mu(A) \cdot \nu(B)$$

is a measure, if the symbols $0 \cdot \infty = 0$ make sense.

Proposition 5.3. *If $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ are measure spaces, then $\big(X \times Y, \mathfrak{Semi}(\mathfrak{S}, \mathfrak{T}), \mu \times \nu\big)$ is also a measure space. And if $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ are σ-finite, then $\big(X \times Y, \mathfrak{Semi}(\mathfrak{S}, \mathfrak{T}), \mu \times \nu\big)$ is also σ-finite.*

Remark 5.4. By Proposition 2.14, the triples

$$(X, \mu^*\text{-}\mathfrak{Meas}, \mu^*) \text{ and } (Y, \nu^*\text{-}\mathfrak{Meas}, \nu^*)$$

are measure spaces. By Proposition 5.3, the triple

$$\big(X \times Y, \mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas}), \mu^* \times \nu^*\big)$$

is also measure space.

Proposition 5.5. *If $A \in \mu^*\text{-}\mathfrak{Meas}$ and $B \in \nu^*\text{-}\mathfrak{Meas}$ with $\mu^*(A) < \infty$ and $\nu^*(B) < \infty$, then*

$$\begin{aligned}(\mu \times \nu)^*(A \times B) &= \big(\mu^* \times \nu^*\big)(A \times B) \\ &= \mu^*(A) \cdot \nu^*(B).\end{aligned}$$

Proposition 5.6. *If $A \in \mu^*\text{-}\mathfrak{Meas}$ and $B \in \nu^*\text{-}\mathfrak{Meas}$, then $A \times B \in (\mu \times \nu)^*\text{-}\mathfrak{Meas}$. That is, $\mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas}) \subseteq (\mu \times \nu)^*\text{-}\mathfrak{Meas}$.*

Proposition 5.7. *If the measure spaces $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ are both σ-finite, then the measure $\mu^* \times \nu^*$ agrees with the measure $(\mu \times \nu)^*$ on the semiring $\mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas})$.*

Proof. We will apply Proposition 2.21.

The measure space $\big(X \times Y, \mathfrak{Semi}(\mathfrak{S}, \mathfrak{T}), \mu \times \nu\big)$ is σ-finite by Proposition 5.3. Since $\mathfrak{S} \subseteq \mu^*\text{-}\mathfrak{Meas}$ and $\mathfrak{T} \subseteq \nu^*\text{-}\mathfrak{Meas}$, it follows that

$$\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T}) \subseteq \mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas}),$$

and by Proposition 5.6, it follows that

$$\mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas}) \subseteq (\mu \times \nu)^*\text{-}\mathfrak{Meas}.$$

By Proposition 5.3, the set function $\mu^* \times \nu^*$ is a measure on $\mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas})$. Further, the measure $\mu^* \times \nu^*$ agrees with the measure $\mu \times \nu$ on $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$ for if $A \in \mathfrak{S}$ and $B \in \mathfrak{T}$, then

$$\begin{aligned}
(\mu^* \times \nu^*)(A \times B) &= \mu^*(A) \cdot \nu^*(B) && \text{Proposition 5.2} \\
&= \mu(A) \cdot \nu(B) && \text{Proposition 2.11} \\
&= (\mu \times \nu)(A \times B) && \text{Proposition 5.2 again}
\end{aligned}$$

Since $\mu^* \times \nu^* = \mu \times \nu$ on $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$, it follows by Proposition 2.21 that $\mu^* \times \nu^* = (\mu \times \nu)^*$ on $\mathfrak{Semi}(\mu^*\text{-}\mathfrak{Meas}, \nu^*\text{-}\mathfrak{Meas})$, completing the proof. ∎

Definition 5.8. Let $\mathfrak{S}$ and $\mathfrak{T}$ be semirings of subsets of respective sets X and Y. The σ-algebra generated by $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$ may be called the **product σ-algebra** of $\mathfrak{S}$ and $\mathfrak{T}$, and will be denoted $\sigma(\mathfrak{S}, \mathfrak{T})$. It is the smallest σ-algebra of $\mathfrak{P}(X \times Y)$ which contains the product semiring $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$. The notation $\mathfrak{S} \otimes \mathfrak{T}$ might also be used, but not here. The notation $\mathfrak{S} \vee \mathfrak{T}$ is compelling.

Notation 5.9. Let us apply Proposition 2.21 to the inclusion

$$\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T}) \subseteq \sigma(\mathfrak{S}, \mathfrak{T}) \subseteq (\mu \times \nu)^*\text{-}\mathfrak{Meas}.$$

The proposition says that if a measure agrees with $\mu \times \nu$ on $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$, then that measure agrees with $(\mu \times \nu)^*$ on $\sigma(\mathfrak{S}, \mathfrak{T})$. We will let $(\mu \times \nu)^*$ denote both the measure $(\mu \times \nu)^*$ on $(\mu \times \nu)^*$-$\mathfrak{Meas}$ and its restriction $(\mu \times \nu)^*_{|\sigma(\mathfrak{S},\mathfrak{T})}$ to $\sigma(\mathfrak{S}, \mathfrak{T})$. We will always make the domain clear.

5.2 Product Integration

Notation 5.10. Let π_1 and π_2 denote the respective **coordinate projections** of $X \times Y$ onto X and Y. That is, $\pi_1(x, y) = x$ and $\pi_2(x, y) = y$.

Definition 5.11. Let $E \subseteq X \times Y$, and let $x \in X$ and $y \in Y$. Define sets

$$E_x \coloneqq \pi_2\big(E \cap \pi_1^{\leftarrow}\{x\}\big) = \{y \in Y : (x, y) \in E\} \subseteq Y,$$

and

$$E^y \coloneqq \pi_1\big(E \cap \pi_2^{\leftarrow}\{y\}\big) = \{x \in X : (x, y) \in E\} \subseteq X.$$

The set E_x is the x-**section** of E , and the set E^y is the y-**section** of E .

Proposition 5.12. *Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces. If $E \subseteq \sigma(\mathfrak{A}, \mathfrak{B})$, then $E_x \in \mathfrak{B}$ for all $x \in X$, and $E^y \in \mathfrak{A}$ for all $y \in Y$.*

Notation 5.13. Whenever X, Y, and Z are sets and $f : X \times Y \to Z$ is a function, we will let f^y denote the map

$$f^y : X \to Z : x \mapsto f(x, y),$$

and we will let f_x denote the map

$$f_x : Y \to Z : y \mapsto f(x, y).$$

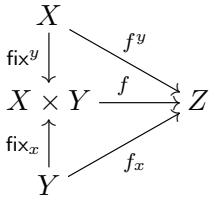
Also, for each $y \in Y$ let fix^y denote the map

$$\mathsf{fix}^y : X \to X \times Y : x \mapsto (x, y)$$

and for each $x \in X$ let fix_x denote the map

$$\mathsf{fix}_x : Y \to X \times Y : y \mapsto (x, y)$$

as illustrated:



Proposition 5.14. *Let $(X, \mathfrak{A})$, $(Y, \mathfrak{B})$, and $(Z, \mathfrak{C})$ be measurable spaces. Let $f : X \times Y \to Z$. If f is $\big(\sigma(\mathfrak{A}, \mathfrak{B}), \mathfrak{C}\big)$-measurable, then f^y is $(\mathfrak{A}, \mathfrak{C})$-measurable for each $y \in Y$, and f_x is $(\mathfrak{B}, \mathfrak{C})$-measurable for each $x \in X$.*

Definition 5.15. Let $E \subseteq X \times Y$. Define the **section functions**

$$\mathsf{Sec}_X E : X \to \mathfrak{P}(Y) : x \mapsto E_x,$$

and

$$\mathsf{Sec}^Y E : Y \to \mathfrak{P}(X) : y \mapsto E^y.$$

See that $\nu^* \circ \mathsf{Sec}_X E : X \to \mathbb{R}$ and $\mu^* \circ \mathsf{Sec}^Y E : Y \to \mathbb{R}$.

Proposition 5.16. *Let $E \in (\mu \times \nu)^*\text{-}\mathfrak{Meas}$ with $(\mu \times \nu)^*(E) < \infty$. For μ^*-almost all $x \in X$, the set E_x is ν^*-measurable. The function $\nu^* \circ \mathsf{Sec}_X E$ defines an integrable function over X with respect to μ, and*

$$(\mu \times \nu)^*(E) = \int_X (\nu^* \circ \mathsf{Sec}_X E)\, d\mu.$$

Likewise, for ν^*-almost all $y \in Y$, the set E^y is μ^*-measurable. The function $\mu^* \circ \mathsf{Sec}^Y E$ defines an integrable function over Y with respect to ν, and

$$(\mu \times \nu)^*(E) = \int_Y (\mu^* \circ \mathsf{Sec}^Y E)\, d\nu.$$

Definition 5.17. Let $f : X \times Y \to \mathbb{R}$. Say that the **iterated integral** $\int_Y \int_X f\, d\mu\, d\nu$ **exists** if, for ν^*-almost all $y \in Y$, the function

$$f^y : X \to \mathbb{R} : x \mapsto f(x, y)$$

is integrable over X with respect to μ, and the function

$$g : Y \to \mathbb{R} : y \mapsto \int_X f^y\, d\mu$$

defines an integrable function over Y with respect to ν. In this case, define

$$\int_Y \int_X f\, d\mu\, d\nu \coloneqq \int_Y \left(\int_X f^y\, d\mu \right) d\nu = \int_Y g\, d\nu$$

Theorem 5.18 (Fubini). *Let $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ both be measure spaces. If $f : X \times Y \to \mathbb{R}$ is a $\mu \times \nu$-integrable function, then both of the iterated integrals $\int_Y \int_X f\, d\mu\, d\nu$ and $\int_X \int_Y f\, d\nu\, d\mu$ exist, and*

$$\int_{X \times Y} f\, d(\mu \times \nu) = \int_Y \int_X f\, d\mu\, d\nu = \int_X \int_Y f\, d\nu\, d\mu.$$

Theorem 5.19 (Tonelli). *Let $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ both be σ-finite measure spaces, and let $f : X \times Y \to \mathbb{R}$ be a $(\mu \times \nu)^*$-measurable function. If one of the iterated integrals $\int_Y \int_X |f|\, d\mu\, d\nu$ or $\int_X \int_Y |f|\, d\nu\, d\mu$ exists, then the function f is $\mu \times \nu$-integrable. Consequently,*

$$\int_{X \times Y} f\, d(\mu \times \nu) = \int_Y \int_X f\, d\mu\, d\nu = \int_X \int_Y f\, d\nu\, d\mu.$$

Remark 5.20. It has been said that, "It is a difficult problem to determine whether a given function $f : X \times Y \to \mathbb{R}$ is $\mu \times \nu$-measurable." Maybe a slight understatement.

5.3 Applications

The following proposition was used in *Elements of Bayesian Statistics*, **1.2.3 Theorem** therein, to show that if a joint measure is absolutely continuous with respect to a product of a marginal measure with a reference measure, then the joint measure is absolutely continuous with respect to the product of the marginal measures. Here, rather than prove absolute continuity, we produce the Radon-Nikodym derivative. Although this proposition might not be needed, we will leave it here for now.

Proposition 5.21. *Let $(X, \mathfrak{S}, \mu)$ and $(Y, \mathfrak{T}, \nu)$ be σ-finite measure spaces, and let $f : X \times Y \to \mathbb{R}$ and $g : X \times Y \to \mathbb{R}$ be non-negative functions, where fg is $(\mu \times \nu)^*$-measurable. If functions f^y and g^y are integrable with respect to μ, and if*

$$\int_Y \left(\int_X f(x,y)\, d\mu(x) \right) \left(\int_X g(x,y)\, d\mu(x) \right) d\nu(y) = 0,$$

then

$$\int_{X \times Y} fg\, d(\mu \times \nu) = 0.$$

Proof. This will be a simple application of Tonelli's. Let functions f^y and g^y be integrable with respect to μ, and let

$$\int_Y \left(\int_X f(x,y)\, d\mu(x) \right) \left(\int_X g(x,y)\, d\mu(x) \right) d\nu(y) = 0.$$

We need to show that $\int_{X\times Y} fg\, d(\mu\times\nu) = 0$.

Let us show that $\int_X f^y g^y\, d\mu = 0$ for ν^*-almost all $y \in Y$. By Proposition 3.37, since the integral of the function

$$Y \to \mathbb{R} : y \mapsto \left(\int_X f(x,y)\, d\mu(x)\right)\left(\int_X g(x,y)\, d\mu(x)\right)$$

with respect to ν equals zero, it follows that the function itself is equal to zero ν^*-almost everywhere on Y, which in turn says that the product of integrals

$$\left(\int_X f(x,y)\, d\mu(x)\right)\left(\int_X g(x,y)\, d\mu(x)\right) \tag{5.1}$$

equals zero ν^*-almost everywhere on Y. The product (5.1) can fail to be zero on a ν^*-null set; let N be the ν^*-null subset of Y on which equality fails. It means that the product (5.1) equals zero everywhere on $Y \setminus N$. Choose one of the factors, say the first, and split $Y \setminus N$ into disjoint subsets on which the first factor is either strictly positive or not; set

$$A = \left\{ y \in Y \setminus N : \int_X f(x,y)\, d\mu(x) > 0 \right\},$$

$$B = \left\{ y \in Y \setminus N : \int_X f(x,y)\, d\mu(x) = 0 \right\}.$$

If $y \in A$, then the second factor $\int_X g(x,y)\, d\mu(x)$ of (5.1) must equal zero, so again by Proposition 3.37, the function $g^y = 0$ μ^*-almost everywhere on X. Likewise, if $y \in B$, then $f^y = 0$ μ^*-almost everywhere on X. This shows that the product of functions $f^y g^y$ equals zero μ^*-almost everywhere on X for all y in the disjoint union $A \cup B = Y \setminus N$. Consequently, $\int_X f^y g^y\, d\mu = 0$ for ν^*-almost all $y \in Y$.

We can say then that the iterated integral $\int_Y \int_X f^y g^y \, d\mu \, d\nu$ exists since for ν^*-almost all $y \in Y$, as we have just shown, the function

$$f^y g^y : X \to \mathbb{R} : x \mapsto f(x,y)g(x,y)$$

is integrable over X with respect to μ having integral zero, and the function

$$Y \to \mathbb{R} : y \mapsto \int_X f^y g^y \, d\mu$$

defines an integrable function over Y with respect to ν because, as we have also just shown, it is the zero function ν^*-almost everywhere. By Tonelli's Theorem 5.19 then, the function fg is $\mu \times \nu$-integrable, and

$$\int_{X\times Y} fg \, d(\mu \times \nu) = \int_Y \int_X f^y g^y \, d\mu \, d\nu = \int_Y 0 \, d\nu = 0,$$

as required. ■

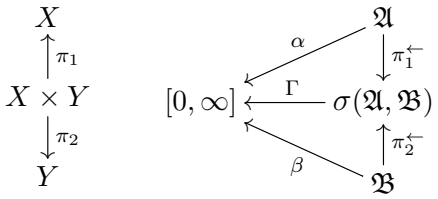
Definition 5.22. Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces. Let Γ be any measure on $\sigma(\mathfrak{A}, \mathfrak{B})$. Define

$$\alpha := \Gamma \circ \pi_1^{\leftarrow} \quad \text{on } \mathfrak{A}$$

and

$$\beta := \Gamma \circ \pi_2^{\leftarrow} \quad \text{on } \mathfrak{B}.$$

Call Γ the **joint** measure. The induced measures α and β are the **marginal** measures, as illustrated:



The marginal measures are determined by the joint. See that

$$\alpha(A) = \Gamma(A \times Y) \quad \text{for all } A \in \mathfrak{A}$$

and

$$\beta(B) = \Gamma(X \times B) \quad \text{for all } B \in \mathfrak{B}$$

since $\pi_1^{\leftarrow}(A) = A \times Y$ and $\pi_2^{\leftarrow}(B) = X \times B$.

In case $(X, \mathfrak{A}, \mu)$ and $(Y, \mathfrak{B}, \nu)$ are already measure spaces, we might refer to the measures μ and ν as **reference** measures, with the understanding that the reference measures are usually not the same as the marginal measures.

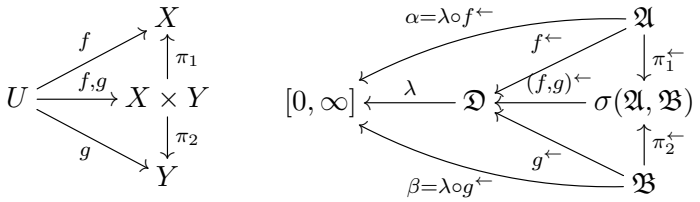
Proposition 5.23. *Let λ be a measure on a σ-algebra $\mathfrak{D}$, and let $(f, g) : U \to X \times Y$ be a $\big(\mathfrak{D}, \sigma(\mathfrak{A}, \mathfrak{B})\big)$-measurable function. Define the joint $\Gamma := \lambda \circ (f, g)^{\leftarrow}$, with marginal $\alpha = \Gamma \circ \pi_1^{\leftarrow}$ and marginal $\beta = \Gamma \circ \pi_2^{\leftarrow}$. Then*

$$\alpha = \lambda \circ f^{\leftarrow}$$

and

$$\beta = \lambda \circ g^{\leftarrow}$$

as illustrated:



Proof. This is just "set theory." For example, while showing $\Gamma \circ \pi_1^{\leftarrow} = \lambda \circ f^{\leftarrow}$, you might use the fact that $\pi_1^{\leftarrow}(S) = S \times Y$ for any $S \subseteq X$. ■

Proposition 5.24. *Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces, and let μ and ν be σ-finite measures on respective σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$. Let Γ be a finite measure on $\sigma(\mathfrak{A}, \mathfrak{B})$. If $\Gamma \ll (\mu \times \nu)^*$ on $\sigma(\mathfrak{A}, \mathfrak{B})$, then $\alpha \ll \mu$ and $\beta \ll \nu$. In particular,*

$$\frac{d\alpha}{d\mu} : X \to \mathbb{R} : x \mapsto \int_Y \frac{d\Gamma}{d(\mu \times \nu)^*}(x, y)\, d\nu(y)$$

and

$$\frac{d\beta}{d\nu} : Y \to \mathbb{R} : y \mapsto \int_X \frac{d\Gamma}{d(\mu \times \nu)^*}(x, y)\, d\mu(x).$$

Proof. Let $f = d\Gamma/d(\mu \times \nu)^*$. We will only show $d\alpha/d\mu : X \to \mathbb{R} : x \mapsto \int_Y f(x, y)\, d\nu(y)$. This is just an application of Fubini's theorem.

The function f is supposed to be integrable over $X \times Y$ with respect to $\mu \times \nu$. It follows by Fubini's Theorem 5.18 that the function

$$f_x : Y \to \mathbb{R} : y \mapsto f(x, y)$$

is integrable over Y with respect to ν for μ^*-almost all $x \in X$, and the function

$$j : X \to \mathbb{R} : x \mapsto \int_Y f_x\, d\nu = \int_Y f(x, y)\, d\nu(y)$$

defines an integrable function over X with respect to μ, with

$$\int_{X \times Y} f\, d(\mu \times \nu) = \int_X j\, d\mu = \int_X \left(\int_Y f(x, y)\, d\nu(y) \right) d\mu(x).$$

For $A \in \mathfrak{A}$, we know

$$\alpha(A) = \Gamma \circ \pi_1^{\leftarrow}(A) = \Gamma(A \times Y).$$

By hypothesis,

$$\Gamma(A \times Y) = \int_{A \times Y} f\, d(\mu \times \nu).$$

By Fubini's,

$$\int_{A\times Y} f\, d(\mu\times\nu) = \int_A j\, d\mu.$$

By transitivity of equality,

$$\alpha(A) = \int_A j\, d\mu \quad \text{for all } A\in\mathfrak{A},$$

and this says $\alpha \ll \mu$ with $j = d\alpha/d\mu$. ■

Proposition 5.25. *Let $(X,\mathfrak{A})$ and $(Y,\mathfrak{B})$ be measurable spaces, and let μ and ν be σ-finite measures on respective σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$. Let Γ be a finite measure on $\sigma(\mathfrak{A},\mathfrak{B})$. If $\Gamma \ll (\mu\times\nu)^*$ on $\sigma(\mathfrak{A},\mathfrak{B})$, then $\Gamma \ll (\alpha\times\nu)^*$, and $\Gamma \ll (\mu\times\beta)^*$. Furthermore, $\Gamma \ll (\alpha\times\beta)^*$. In particular,*

$$\frac{d\Gamma}{d(\alpha\times\nu)^*} : X\times Y\to\mathbb{R} : (x,y)\mapsto \frac{\dfrac{d\Gamma}{d(\mu\times\nu)^*}(x,y)}{\dfrac{d\alpha}{d\mu}(x)},$$

$$\frac{d\Gamma}{d(\mu\times\beta)^*} : X\times Y\to\mathbb{R} : (x,y)\mapsto \frac{\dfrac{d\Gamma}{d(\mu\times\nu)^*}(x,y)}{\dfrac{d\beta}{d\nu}(y)},$$

and

$$\frac{d\Gamma}{d(\alpha\times\beta)^*} : X\times Y\to\mathbb{R} : (x,y)\mapsto \frac{\dfrac{d\Gamma}{d(\mu\times\nu)^*}(x,y)}{\dfrac{d\alpha}{d\mu}(x)\cdot\dfrac{d\beta}{d\nu}(y)}.$$

Proof. Let $\Gamma \ll (\mu \times \nu)^*$. Let $f = d\Gamma/d(\mu \times \nu)^*$ and let

$$j : X \to \mathbb{R} : x \mapsto \int_Y f_x \, d\nu$$

and

$$k : Y \to \mathbb{R} : y \mapsto \int_X f^y \, d\mu,$$

and define

$$g : X \times Y \to \mathbb{R} : (x, y) \mapsto \begin{cases} \dfrac{f(x,y)}{j(x)} & \text{if } j(x) > 0 \\ 1 & \text{if } j(x) = 0 \end{cases}$$

and

$$h : X \times Y \to \mathbb{R} : (x, y) \mapsto \begin{cases} \dfrac{f(x,y)}{j(x)k(y)} & \text{if } j(x)k(y) > 0 \\ 1 & \text{if } j(x)k(y) = 0 \end{cases}$$

The claim is that $g = d\Gamma/d(\alpha \times \nu)^*$ and $h = d\Gamma/d(\alpha \times \beta)^*$.

Let us now show that $\Gamma \ll (\alpha \times \nu)^*$. It is sufficient to show that $g = d\Gamma/d(\alpha \times \nu)^*$, meaning

$$\Gamma(C) = \int_C g \, d(\alpha \times \nu) \quad \text{for all } C \in \sigma(\mathfrak{A}, \mathfrak{B}).$$

In order to apply Tonelli's Theorem 5.19 to show that the function g is integrable with respect to $\alpha \times \nu$, we first show that the iterated integral $\int_X \int_Y g \, d\nu \, d\alpha$ exists.

We have assumed $f = d\Gamma/d(\mu \times \nu)^*$, which implies that we have assumed the function f is integrable with respect to $\mu \times \nu$. By Fubini's Theorem 5.18 then, the corresponding iterated integral $\int_X \int_Y f \, d\nu \, d\mu$ exists. This implies that the function

$$f_x : Y \to \mathbb{R} : y \mapsto f(x, y)$$

is integrable over Y with respect to ν for μ^*-almost all $x \in X$, and the function $x \mapsto \int_Y f_x \, d\nu$ defines an integrable function over X with respect to μ. Let us carry this a little further. Since $\Gamma \ll (\mu \times \nu)^*$, it follows by Proposition 5.24, that $\alpha \ll \mu$ on $\mathfrak{A}$. We can then apply Proposition 4.25 which says that if $\mu^*(A) = 0$, then $\alpha^*(A) = 0$ for any $A \in X$. It means the function f_x is integrable over Y with respect to ν for α^*-almost all $x \in X$. Since the function $Y \to \mathbb{R} : y \mapsto 1/j(x)$ is constant, it follows that the function

$$\frac{f_x}{j(x)} : Y \to \mathbb{R} : y \mapsto \frac{f(x,y)}{j(x)}$$

is also integrable over Y with respect to ν for α^*-almost all $x \in X$. See that the function

$$X \to \mathbb{R} : x \mapsto \int_Y \frac{f_x}{j(x)} \, d\nu$$

defines an integrable function over X with respect to α; in fact,

$$\int_Y \frac{f_x}{j(x)} \, d\nu = 1,$$

and constant functions are integrable with respect to finite measures.

Consequently, for $C \in \sigma(\mathfrak{A}, \mathfrak{B})$, the function

$$Y \to \mathbb{R} : y \mapsto \frac{\chi_C(x,y) f_x(y)}{j(x)}$$

is also integrable over Y with respect to ν for α^*-almost all $x \in X$, and the function

$$X \to \mathbb{R} : x \mapsto \int_Y \frac{\chi_C(x,y) f_x(y)}{j(x)} \, d\nu(y)$$

defines an integrable function over X with respect to α. Applying Tonelli's Theorem 5.19 then,

$$\begin{aligned}\int_{X\times Y} \chi_C g \, d(\alpha\times\nu) &= \int_{X\times Y} \frac{\chi_C(x,y)f_x(y)}{j(x)}\, d(\alpha\times\nu)(x,y)\\ &= \int_Y \int_X \frac{\chi_C(x,y)f_x(y)}{j(x)}\, d\alpha(x)\, d\nu(y).\end{aligned}$$

By Proposition 4.29, we can change measures, after $d\alpha = j d\mu$:

$$\begin{aligned}\int_X \frac{\chi_C(x,y)f_x(y)}{j(x)}\, d\alpha(x) &= \int_X \frac{\chi_C(x,y)f_x(y)}{j(x)} \cdot j(x)\, d\mu(x)\\ &= \int_X \chi_C(x,y)f_x(y)\, d\mu(x).\end{aligned}$$

Using these last two strings of equalities:

$$\begin{aligned}\int_C g\, d(\alpha\times\nu) &= \int_{X\times Y} \chi_C g\, d(\alpha\times\nu)\\ &= \int_Y\int_X \chi_C(x,y)f_x(y)\, d\mu(x)\, d\nu(y)\\ &= \int_C f(x,y)\, d(\mu\times\nu)\\ &= \Gamma(C),\end{aligned}$$

as required.

Let us sketch the proof that $h = d\Gamma/d(\alpha\times\beta)^*$. We have shown that $\Gamma \ll (\mu\times\nu)^*$ implies $\Gamma \ll (\alpha\times\nu)^*$. Simply use the symmetric proof to show $\Gamma \ll (\alpha\times\nu)^*$ implies $h = d\Gamma/d(\alpha\times\beta)^*$. Use the fact shown earlier that $g \in L_1(\alpha)$ and so

$$\int_X g(x,y)\, d\alpha(x) = \int_X \frac{f(x,y)}{j(x)}\, d\alpha(x) = \int_X \frac{f(x,y)}{j(x)} j(x)\, d\mu(x) = k(y)$$

to get

$$\int_X \frac{g^y}{k(y)}\, d\alpha = 1$$

in order to show that the function

$$Y \to \mathbb{R} : y \mapsto \int_X \frac{g^y}{k(y)}\, d\alpha$$

defines an integrable function over Y with respect to the finite measure β. And, for future reference:

$$\int_X \frac{g^y}{k(y)}\, d\alpha = \int_X \frac{\dfrac{d\Gamma}{d(\mu\times\nu)^*}(x,y)}{\dfrac{d\alpha}{d\mu}(x)\cdot\dfrac{d\beta}{d\nu}(y)}\, d\alpha(x).$$

■

Proposition 5.26. *Let $(X,\mathfrak{A})$ and $(Y,\mathfrak{B})$ be measurable spaces, and let μ and ν be σ-finite measures on respective σ-algebras $\mathfrak{A}$ and $\mathfrak{B}$. Let Γ be a finite measure on $\sigma(\mathfrak{A},\mathfrak{B})$. If $\Gamma \ll (\alpha\times\nu)^*$, then $\beta \ll \nu$ and $\Gamma \ll (\alpha\times\beta)^*$. In particular,*

$$\frac{d\beta}{d\nu} : Y \mapsto \mathbb{R} : y \mapsto \int_X \frac{d\Gamma}{d(\alpha\times\nu)^*}(x,y)\, d\alpha(x),$$

and

$$\frac{d\Gamma}{d(\alpha\times\beta)^*} : X\times Y \to \mathbb{R} : (x,y) \mapsto \frac{\dfrac{d\Gamma}{d(\alpha\times\nu)^*}(x,y)}{\dfrac{d\beta}{d\nu}(y)}.$$

Likewise, if $\Gamma \ll (\mu\times\beta)^*$, then $\alpha \ll \mu$ and $\Gamma \ll (\alpha\times\beta)^*$. In particular,

$$\frac{d\alpha}{d\mu} : X \mapsto \mathbb{R} : x \mapsto \int_Y \frac{d\Gamma}{d(\mu\times\beta)^*}(x,y)\, d\beta(y),$$

and

$$\frac{d\Gamma}{d(\alpha\times\beta)^*} : X\times Y\to\mathbb{R} : (x,y)\mapsto \frac{\dfrac{d\Gamma}{d(\mu\times\beta)^*}(x,y)}{\dfrac{d\alpha}{d\mu}(x)}.$$

Proof. This is nothing but an application of Proposition 5.24 and Proposition 5.25. First set $\mu=\alpha$ and apply them, then set $\nu=\beta$ and apply them again, noting that α and β are both finite and therefore σ-finite. ■

Proposition 5.27. *Let $(X,\mathfrak{A})$ and $(Y,\mathfrak{B})$ be measurable spaces, and let Γ be a finite measure on $\sigma(\mathfrak{A},\mathfrak{B})$. If $\Gamma\ll(\alpha\times\beta)^*$, then*

$$X\to\mathbb{R} : x\mapsto \int_Y \frac{d\Gamma}{d(\alpha\times\beta)^*}(x,y)\,d\beta(y)=1 \quad \alpha^*\text{-a.e. on } X,$$

and

$$Y\to\mathbb{R} : y\mapsto \int_X \frac{d\Gamma}{d(\alpha\times\beta)^*}(x,y)\,d\alpha(x)=1 \quad \beta^*\text{-a.e. on } Y.$$

Proof. We will prove the first displayed equality. Let $\Gamma\ll(\alpha\times\beta)^*$. With Γ finite, it follows that $\alpha:=\Gamma\circ\pi_1^{\leftarrow}$ on $\mathfrak{A}$ and $\beta:=\Gamma\circ\pi_2^{\leftarrow}$ on $\mathfrak{B}$ are finite, therefore σ-finite. If we now set $\mu=\alpha$ and set $\nu=\beta$ and apply Proposition 5.24, then we get:

$$\frac{d\alpha}{d\alpha}(x)=\int_Y \frac{d\Gamma}{d(\alpha\times\beta)^*}(x,y)\,d\beta(y),$$

or equivalently,

$$\alpha(A)=\int_A\left(\int_Y \frac{d\Gamma}{d(\alpha\times\beta)^*}(x,y)\,d\beta(y)\right)d\alpha(x) \quad \text{for all } A\in\mathfrak{A}.$$

But also

$$\alpha(A) = \int_A 1 \, d\alpha(x) \quad \text{for all } A \in \mathfrak{A}.$$

By the uniqueness of the Radon-Nikodym derivative, which in this case is a class in $L_1(X, \mathfrak{A}, \alpha)$, it follows that

$$\int_Y \frac{d\Gamma}{d(\alpha \times \beta)^*}(x, y) \, d\beta(y) = 1 \quad \alpha^*\text{-almost everywhere,}$$

as required. ■

6 Probability, Independence, and Sampling

6.1 Probability Spaces and Distributions

Definition 6.1. A **probability space** is a measure space $(X, \mathfrak{A}, \mu)$ such that $\mathfrak{A}$ is a σ-algebra of subsets of a set X with $\mu(X) = 1$. Any measure ν on a measurable space $(X, \mathfrak{A})$ for which $\nu(X) = 1$ is called a **probability measure**, and if $f : X \to \mathbb{R}^n$ is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}^n})$-measurable function, then f might be called a **random variable**.

Remark 6.2. Why change structure from measure spaces $(X, \mathfrak{S}, \mu)$, where $\mathfrak{S}$ is only assumed to be a semiring, to $(X, \mathfrak{A}, \mu)$, where $\mathfrak{A}$ is assumed to be a σ-algebra? One reason might be the use of *density* functions, which represent a Radon-Nikodym derivative. Recall that the Radon-Nikodym Theorem 4.27 applies to measures on a measurable space $(X, \mathfrak{A})$, in which case $\mathfrak{A}$ is supposed to be a σ-algebra. Density functions are described further in Notation 6.4. A consequence of this change of structure is that the measure $\mu \times \nu$ on the semiring $\mathfrak{Semi}(\mathfrak{A}, \mathfrak{B})$ need no longer be a measure on $\sigma(\mathfrak{A}, \mathfrak{B})$, so we will use its unique extension $(\mu \times \nu)^*$.

Definition 6.3. Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $(Y, \mathfrak{B})$ be a measurable space, so $\mathfrak{B}$ is supposed to be a σ-algebra. If a function $f : X \to Y$ is $(\mathfrak{A}, \mathfrak{B})$-measurable, where $f^{\leftarrow} : \mathfrak{B} \to \mathfrak{A}$, then

the induced probability measure $\mu \circ f^{\leftarrow}$ on $\mathfrak{B}$, defined by

$$\mu \circ f^{\leftarrow} : \mathfrak{B} \to [0,1] : B \mapsto \mu\big(f^{\leftarrow}(B)\big),$$

and illustrated

$$X \xrightarrow{\ f\ } Y \qquad [0,1] \xleftarrow[\mu]{} \mathfrak{A} \xleftarrow[f^{\leftarrow}]{} \mathfrak{B}, \quad \mu \circ f^{\leftarrow} : \mathfrak{B} \to [0,1]$$

is called the **probability distribution** of f on $\mathfrak{B}$, or simply the *probability distribution* of f.

Notation 6.4. Probability distributions are probabilities induced on a σ-algebra of the codomain of a measurable function. The notation

$$f \sim \mu \circ f^{\leftarrow}$$

is used to denote the map from the measurable function to the probability distribution, although seldom so explicitly. For example, suppose an author writes

$$f \sim \text{Normal}(m, s),$$

or maybe $f \sim \mathscr{N}(m, s)$. There is an underlying probability space $(X, \mathfrak{A}, \mu)$, where $f : X \to \mathbb{R}$ is $\big(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}}\big)$-measurable, as illustrated:

$$X \xrightarrow{\ f\ } \mathbb{R} \qquad [0,1] \xleftarrow[\mu]{} \mathfrak{A} \xleftarrow[f^{\leftarrow}]{} \mathfrak{Bor}_{\mathbb{R}}, \quad \mu \circ f^{\leftarrow} = \text{Normal}(m,s) : \mathfrak{Bor}_{\mathbb{R}} \to [0,1]$$

See that $f \sim \mu \circ f^{\leftarrow}$. That is, $\mu \circ f^{\leftarrow}$ is the probability distribution of f on $\mathfrak{Bor}_{\mathbb{R}}$. By writing $f \sim \text{Normal}(m, s)$, the author has indicated

that the measures $\mu \circ f^{\leftarrow}$ and Normal(m, s) are equal, which means

$$\begin{aligned}\mu \circ f^{\leftarrow}(A) &= \text{Normal}(m, s)(A) \\ &\coloneqq \int_A g(t)\, d\lambda(t) \quad \text{for all } A \in \mathfrak{Bor}_{\mathbb{R}},\end{aligned}$$

where

$$g : \mathbb{R} \to \mathbb{R} : t \mapsto \frac{1}{s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-m}{s}\right)^2\right\},$$

and λ is Lebesgue measure. The function $g \in \mathscr{L}_1(\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}, \lambda)$ in this case represents the Radon-Nikodym derivative of the probability measure $\lambda \circ f^{\leftarrow}$ with respect to Lebesgue measure λ on $\mathfrak{Bor}_{\mathbb{R}}$. Should an author write

$$\text{Normal}(f \mid m, s),$$

they are likely referring to the function g itself, and might call it the **density** of f.

If an author writes $\int g(t)\, dt$, then we assume the 'dt' means with respect to Lebesgue measure. If instead there is a summation sign and they use the term 'probability function,' or use the word 'discrete,' then they likely mean integration with respect to counting measure. In any case, with a density function it is always the Lebesgue integral with respect to some measure.

Remark 6.5. To be sure:

- A *density* is an integrable function that represents a Radon-Nikodym derivative of a finite measure with respect to a σ-finite measure.
- A *probability distribution* is a probability measure induced on a σ-algebra of the codomain of a measurable function.

There is no intelligible sense in which the terms *density* and *probability distribution* are interchangeable. Still, there are some authors that will confuse them.

6.2 Independence

Definition 6.6. Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $\mathfrak{T}$ and $\mathfrak{U}$ be subsets of the semiring $\mathfrak{S}$. Then $\mathfrak{T}$ and $\mathfrak{U}$ are **independent collections** of sets with respect to the measure μ if

$$\mu(A \cap B) = \mu(A) \cdot \mu(B) \quad \text{for all } A \in \mathfrak{T} \text{ and for all } B \in \mathfrak{U}.$$

Proposition 6.7. *If $\mathfrak{T}$ and $\mathfrak{U}$ are independent collections with $\mathfrak{V} \subseteq \mathfrak{T}$ and $\mathfrak{W} \subseteq \mathfrak{U}$, then $\mathfrak{V}$ and $\mathfrak{W}$ are independent collections.*

Definition 6.8. Let $f : (X, \mathfrak{A}) \to (Y, \mathfrak{B})$ and $g : (X, \mathfrak{A}) \to (Z, \mathfrak{C})$ be measurable functions. Then f and g are **independent functions** with respect to a measure μ on $\mathfrak{A}$ if the σ-algebras $f^{\leftarrow}(\mathfrak{B})$ and $g^{\leftarrow}(\mathfrak{C})$ are independent with respect to μ, where $f^{\leftarrow} : \mathfrak{B} \to \mathfrak{A}$ and $g^{\leftarrow} : \mathfrak{C} \to \mathfrak{A}$. It means

$$\mu\big(f^{\leftarrow}(A) \cap g^{\leftarrow}(B)\big) = \mu\big(f^{\leftarrow}(A)\big) \cdot \mu\big(g^{\leftarrow}(B)\big) \quad \text{for all } A \in \mathfrak{A} \text{ and } B \in \mathfrak{B}.$$

Proposition 6.9. *If functions $h : (X, \mathfrak{A}) \to (Y, \mathfrak{B})$ and $k : (X, \mathfrak{A}) \to (Y, \mathfrak{B})$ are independent, and $g : (Y, \mathfrak{B}) \to (Z, \mathfrak{C})$ is measurable, then $g \circ h$ and $g \circ k$ are independent.*

Definition 6.10. Let $(X, \mathfrak{S}, \mu)$ be a measure space, and let $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ be classes in $L_p(X, \mathfrak{S}, \mu)$. Then $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ are **independent classes** with respect to μ^* if the $(\mu^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable functions f and g are independent with respect to μ^*. Proposition 3.10 says that the independence of classes $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ does not depend

on the respective chosen representative functions f and g, meaning if $h \in \overline{f}^{\mu}$ and $k \in \overline{g}^{\mu}$, then f and g are independent if and only if h and k are independent.

Definition 6.11. For a class $\overline{f}^{\mu}$ in $L_1(X, \mathfrak{S}, \mu)$, the **expected value** $\mathsf{E}\,\overline{f}^{\mu}$ of the class $\overline{f}^{\mu}$ is $\int f\,d\mu$. That is, the operator E is the linear functional defined by

$$\mathsf{E} : L_1(X, \mathfrak{S}, \mu) \to \mathbb{R} : \overline{f}^{\mu} \mapsto \int_X f\,d\mu.$$

Proposition 6.12. *If classes $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ in $L_1(X, \mathfrak{S}, \mu)$ are independent with respect to μ^*, and if $\overline{f}^{\mu}\overline{g}^{\mu} \in L_1(X, \mathfrak{S}, \mu)$, then*

$$\mathsf{E}\left(\overline{f}^{\mu}\overline{g}^{\mu}\right) = \left(\mathsf{E}\,\overline{f}^{\mu}\right)\left(\mathsf{E}\,\overline{g}^{\mu}\right).$$

Proof. This is nothing more imaginative than wading through the definitions of independence, and showing that the equality holds for step functions. Then use the continuity of the integral by recalling that the equivalence classes of step functions are norm dense in $L_1(X, \mathfrak{S}, \mu)$. For example, let $f = a\chi_A + b\chi_B$ and $g = c\chi_C + d\chi_D$ be independent with respect to μ^*, where we suppose A, B, C, and D to be disjoint and μ^*-measurable sets of finite measure. Then

$$\begin{aligned} fg &= (a\chi_A + b\chi_B)(c\chi_C + d\chi_D) \\ &= ac\chi_{A\cap C} + ad\chi_{A\cap D} + bc\chi_{B\cap C} + bd\chi_{B\cap D}, \end{aligned}$$

and so

$$\begin{aligned} \int fg &= ac\mu^*(A\cap C) + ad\mu^*(A\cap D) + bc\mu^*(B\cap C) + bd\mu^*(B\cap D) \\ &= ac\mu^*(A)\mu^*(C) + ad\mu^*(A)\mu^*(D) + bc\mu^*(B)\mu^*(C) + bd\mu^*(B)\mu^*(D) \\ &= \big(a\mu^*(A) + b\mu^*(B)\big)\big(c\mu^*(C) + d\mu^*(D)\big) \\ &= \int f \int g. \end{aligned}$$

■

6.3 Strong Law of Large Numbers

Theorem 6.13 (Strong Law of Large Numbers). *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\{f, f_1, f_2, \ldots\}$ be a collection of functions in $\mathscr{L}_1(X, \mathfrak{A}, \mu)$. Define the subset*

$$A \coloneqq \left\{ x \in X : \frac{(f_1 + \cdots + f_n)(x)}{n} \to \int_X f \, d\mu \right\}.$$

If the f_i are independent and all have the same probability distribution as f, then $\mu(A) = 1$.

6.4 Sampling

There are some terms we will not even try to define for now. Terms like *draw*, or *sample*, or even *random*. Especially *random*.

Proposition 6.14. *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $(Y, \mathfrak{B}, \nu)$ be σ-finite. Also let $\{h_i : X \to Y\}$ be a collection of $(\mathfrak{A}, \mathfrak{B})$-measurable functions. If the h_i are independent, and if the probability distributions of the h_i all have the same density function f with respect to the measure ν, and if $g \in \mathscr{L}_1(Y, \mathfrak{B}, \mu \circ h_i^{\leftarrow})$, then the subset of X defined by*

$$A \coloneqq \left\{ x \in X : \frac{1}{n} \sum_{i=1}^{n} \big(g \circ h_i\big)(x) \to \int_Y gf \, d\nu \right\}$$

has $\mu(A) = 1$.

Proof. Let the h_i be independent, and let the probability distributions $\mu \circ h_i^{\leftarrow}$ of the h_i all have the same density function f with respect to ν on $\mathfrak{B}$, and let $g \in \mathscr{L}_1(Y, \mathfrak{B}, \mu \circ h_i^{\leftarrow})$. We will use the strong law of large numbers to show that the μ-measure of the subset

$$\left\{ x \in X : \frac{1}{n} \sum_{i=1}^{n} \big(g \circ h_i\big)(x) \to \int_X (g \circ h_1)\, d\mu \right\}$$

is 1, and to complete the proof we will show that

$$\int_X (g \circ h_1)\, d\mu = \int_Y gf\, d\nu.$$

For the probability distributions $\mu \circ h_i^{\leftarrow}$ of the h_i to have the same density function f in $\mathscr{L}_1(Y, \mathfrak{B}, \nu)$ with respect to ν on $\mathfrak{B}$ means

$$(\mu \circ h_i^{\leftarrow})(B) = \int_B f\, d\nu \quad \text{for all } B \in \mathfrak{B},$$

as somewhat illustrated:

$$X \xrightarrow{h_i} Y \xrightarrow{f} \mathbb{R} \qquad [0,1] \overset{\overset{\mu \circ h_i^{\leftarrow}}{\curvearrowleft}}{\underset{\mu}{\leftarrow}} \mathfrak{A} \underset{h_i^{\leftarrow}}{\leftarrow} \mathfrak{B}$$

The collection $\{g \circ h_i\}$ of functions are independent with respect to μ by Proposition 6.9. By hypothesis $g \in \mathscr{L}_1(Y, \mathfrak{B}, \mu \circ h_i^{\leftarrow})$, and so by Proposition 4.13, each $g \circ h_i \in \mathscr{L}_1(X, \mathfrak{A}, \mu)$. To show that the collection $\{g \circ h_i\}$ of functions all have the same probability distribution as $g \circ h_1$, see first that the $\mu \circ h_i^{\leftarrow} = \mu \circ h_1^{\leftarrow}$ since by hypothesis the $\mu \circ h_i^{\leftarrow}$ all have the same density function with respect to λ:

$$\mu \circ h_i^{\leftarrow}(B) = \int_B f\, d\lambda = \mu \circ h_1^{\leftarrow}(B) \quad \text{for all } B \in \mathfrak{Bor}_{\mathbb{R}}.$$

See next that the $\{g \circ h_i\}$ all have the same probability distribution as $g \circ h_1$ since

$$\begin{aligned}\mu \circ (g \circ h_i)^{\leftarrow} &= \mu \circ (h_i^{\leftarrow} \circ g^{\leftarrow}) \\ &= (\mu \circ h_i^{\leftarrow}) \circ g^{\leftarrow} \\ &= (\mu \circ h_1^{\leftarrow}) \circ g^{\leftarrow} \\ &= \mu \circ (g \circ h_1)^{\leftarrow}.\end{aligned}$$

Since the collection $\{g \circ h_i\}$ of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions are independent with respect to the measure μ, and they all have the same probability distribution as $g \circ h_1$, it follows by the strong law of large numbers, Theorem 6.13, that the subset

$$A \coloneqq \left\{ x \in X : \frac{1}{n} \sum_{i=1}^{n} \big(g \circ h_i\big)(x) \to \int_X (g \circ h_1)\, d\mu \right\}$$

of X has $\mu(A) = 1$.

Let us now show that

$$\int_X (g \circ h_1)\, d\mu = \int_Y g\, d(\mu \circ h_1^{\leftarrow}).$$

The hypothesis that $g \in \mathscr{L}_1(\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}, \mu \circ h_i^{\leftarrow})$ means that Proposition 4.13 can be applied, which says that since h_1 is $(\mathfrak{A}, \mathfrak{B})$-measurable and g is integrable with respect to $\mu \circ h_1^{\leftarrow}$, it follows that $g \circ h_1$ is integrable with respect to μ, and $\int_X (g \circ h_1)\, d\mu = \int_Y g\, d(\mu \circ h_1^{\leftarrow})$. The picture there was

$$\begin{array}{ccccc} (X, \mathfrak{A}, \mu) & & \xrightarrow{\quad g \circ h_1 \quad} & & \\ & \xrightarrow[h_1]{} & (Y, \mathfrak{B}, \mu \circ h_1^{\leftarrow}) & \xrightarrow{g} & (\mathbb{R}, \mathfrak{Bor}_{\mathbb{R}}), \\ h_1^{\leftarrow}(B) & & B & & \end{array}$$

with

$$\int_{h_1^{\leftarrow}(B)} (g \circ h_1)\, d\mu = \int_B g\, d(\mu \circ h_1^{\leftarrow}).$$

Finally, let us show that

$$\int_Y g\, d(\mu \circ h_1^{\leftarrow}) = \int_Y gf\, d\nu.$$

Again, by hypothesis, the function g is integrable with respect to $\mu \circ h_1^{\leftarrow}$. Further, $f = d(\mu \circ h_1^{\leftarrow})/d\nu$, with $\mu \circ h_1^{\leftarrow}$ finite and ν σ-finite, so Proposition 4.29 (2) says that the function gf is integrable with respect to ν, and that $\int_Y g\, d(\mu \circ h_1^{\leftarrow}) = \int_Y gf\, d\nu$, completing the proof. ■

7 Conditional Mathematical Expectation

CONDITIONAL mathematical expectation is not difficult to define, but there *seems* to be a subtlety in the proof of its existence, especially when it comes to applying the Radon-Nikodym theorem.

We are supposing that the '$\mathfrak{A}$' in any probability space $(X, \mathfrak{A}, \mu)$ is a σ-algebra, as opposed to merely a semi-ring.

Definition 7.1 (Conditional Mathematical Expectation). Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\overline{f} \in L_1(X, \mathfrak{A}, \mu)$. The **conditional mathematical expectation** $\mathscr{E}_{\mathfrak{B}}\,\overline{f}$ of the class $\overline{f}$ with respect to the σ-algebra $\mathfrak{B}$, or simply the *conditional expectation* of $\overline{f}$ with respect to $\mathfrak{B}$, is the unique class in $L_1(X, \mathfrak{A}, \mu)$ from which any representative function $g \in \mathscr{L}_1(X, \mathfrak{A}, \mu)$ satisfies the equation

$$\int_B f\,d\mu = \int_B g\,d\mu \quad \text{for all } B \in \mathfrak{B}. \tag{7.1}$$

Such a representative function g from the class $\mathscr{E}_{\mathfrak{B}}\,\overline{f}$ can be chosen to be $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable.

Theorem 7.2. *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. For every σ-subalgebra $\mathfrak{B}$ of $\mathfrak{A}$ there exists a unique class $\overline{g}^{\mu_{|\mathfrak{B}}} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ such that*

$$\int_B f\,d\mu = \int_B g\,d\mu_{|\mathfrak{B}} \quad \textit{for all } B \in \mathfrak{B}, \tag{7.2}$$

where we may suppose the representative function $g \in \mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ *to be* $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$*-measurable.*

Proof. The existence of such a $\overline{g}^{\mu_{|\mathfrak{B}}}$ is an application of the Radon-Nikodym theorem, and to see this, let $\nu_{|\mathfrak{B}}$ denote the restriction of the indefinite integral ν of $\overline{f}^{\mu}$ to the σ-subalgebra $\mathfrak{B}$, which means that $\nu_{|\mathfrak{B}} : \mathfrak{B} \to \mathbb{R}$ is defined by

$$\nu_{|\mathfrak{B}}(B) = \int_B f \, d\mu \quad \text{for all } B \in \mathfrak{B}.$$

The set function $\nu_{|\mathfrak{B}}$ is a finite signed measure that is absolutely continuous with respect to the σ-finite measure $\mu_{|\mathfrak{B}}$ on $\mathfrak{B}$, and so by the Radon-Nikodym Theorem 4.27 there is a unique class $\overline{g}^{\mu_{|\mathfrak{B}}}$ in $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ such that

$$\nu_{|\mathfrak{B}}(B) = \int_B g \, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B},$$

where the representative function g may be taken to be $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, which completes the proof. ■

Definition 7.3. Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. By Theorem 7.2, for each σ-subalgebra $\mathfrak{B}$ of $\mathfrak{A}$ there is a unique class, which may be denoted by $E_{\mathfrak{B}} \overline{f}^{\mu}$ or $E\big(\overline{f}^{\mu} \mid \mathfrak{B}\big)$, in $L_1\big(X, \mathfrak{B}, \mu_{|\mathfrak{B}}\big)$ such that

$$\int_B f \, d\mu = \int_B E_{\mathfrak{B}} \overline{f}^{\mu} \, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B}. \tag{7.3}$$

The class $E_{\mathfrak{B}} \overline{f}^{\mu}$, having a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative, is the **conditional expectation** of $\overline{f}^{\mu}$ in $L_1\big(X, \mathfrak{B}, \mu_{|\mathfrak{B}}\big)$ with respect to the σ-algebra $\mathfrak{B}$. This defines an operator:

$$E_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1\big(X, \mathfrak{B}, \mu_{|\mathfrak{B}}\big) : \overline{f}^{\mu} \mapsto E_{\mathfrak{B}} \overline{f}^{\mu}.$$

The fact that

$$\int_X^{\bar{}} \overline{f}^{\mu}\, d\mu = \int_X f\, d\mu = \int_X^{\bar{}} E_{\mathfrak{B}}\, \overline{f}^{\mu}\, d\mu_{|\mathfrak{B}}$$

implies that the operator $E_{\mathfrak{B}}$ has norm 1.

Notation 7.4. Let $(X, \mathfrak{A}, \mu)$ be a probability space. Let $(Y, \mathfrak{B})$ be a measurable space, and let $g : X \to Y$ be any function, where $g^{\leftarrow} : \mathfrak{B} \to \mathfrak{P}(X)$, with the qualification that in case no σ-algebra $\mathfrak{B}$ is supplied, we take $\mathfrak{B}$ to be the powerset $\mathfrak{P}(Y)$ of Y. Let $\mathfrak{A}^{(g)}$ denote the σ-subalgebra of $\mathfrak{B}$ consisting of those sets in $\mathfrak{B}$ whose preimages under g lie in $\mathfrak{A}$:

$$\mathfrak{A}^{(g)} = \{B \in \mathfrak{B} : g^{\leftarrow}(B) \in \mathfrak{A}\} \subseteq \mathfrak{B}.$$

The σ-subalgebra

$$\mathfrak{G} \coloneqq g^{\leftarrow}(\mathfrak{B}) \cap \mathfrak{A}$$

is comprised of the preimages under g of the sets in $\mathfrak{A}^{(g)}$:

$$\mathfrak{G} = \left\{g^{\leftarrow}(B) : B \in \mathfrak{A}^{(g)}\right\} \subseteq \mathfrak{A}.$$

Consequently, the function g is $\left(\mathfrak{G}, \mathfrak{A}^{(g)}\right)$-measurable. Furthermore, the function defined by

$$g^{\leftarrow} : \mathfrak{A}^{(g)} \twoheadrightarrow \mathfrak{G} : C \mapsto g^{\leftarrow}(C)$$

is surjective.

Definition 7.5. Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. Let $(Y, \mathfrak{B})$ be a measurable space, and let $g : X \to Y$ be any function, where $g^{\leftarrow} : \mathfrak{B} \to \mathfrak{P}(X)$. Temporarily define

$$\mathfrak{G} \coloneqq g^{\leftarrow}(\mathfrak{B}) \cap \mathfrak{A}.$$

Then the class $E\big(\overline{f}^{\mu}\,|\,\mathfrak{G}\big)$ in $L_1\big(X,\mathfrak{G},\mu_{|\mathfrak{G}}\big)$, which may also be denoted by $E\big(\overline{f}^{\mu}\,|\,g\big)$, is the *conditional mathematical expectation of* $\overline{f}^{\mu}$ *given the function* g, or simply the *conditional expectation* of $\overline{f}^{\mu}$ given g. This defines an operator:

$$E_{\mathfrak{G}} : L_1(X,\mathfrak{A},\mu) \to L_1\big(X,\mathfrak{G},\mu_{|\mathfrak{G}}\big) : \overline{f}^{\mu} \mapsto E\big(\overline{f}^{\mu}\,|\,g\big).$$

We use this definition of $E\big(\overline{f}^{\mu}\,|\,g\big)$ because it is precisely the definition of conditional mathematical expectation as set down in Kolmogorov [4].

In what follows, the fact that $(\mu_{|\mathfrak{G}})\circ g^{\leftarrow} = (\mu\circ g^{\leftarrow})_{|\mathfrak{A}^{(g)}}$ will be used. By Proposition 4.34, there is a unique class

$$r + \mathscr{N}_1\big(Y,\mathfrak{A}^{(g)},(\mu\circ g^{\leftarrow})_{|\mathfrak{A}^{(g)}}\big) \;\in\; L_1\big(Y,\mathfrak{A}^{(g)},(\mu\circ g^{\leftarrow})_{|\mathfrak{A}^{(g)}}\big)$$

with

$$E\big(\overline{f}^{\mu}\,|\,g\big) = r\circ g + \mathscr{N}_1\big(X,\mathfrak{G},\mu_{|\mathfrak{G}}\big) = \overline{r\circ g}^{\mu_{|\mathfrak{G}}}$$

and with the representative r being an $(\mathfrak{A}^{(g)},\mathfrak{Bor}_{\mathbb{R}})$-measurable function, as illustrated in the following diagram, letting the function e be any representative of the class $E\big(\overline{f}^{\mu}\,|\,g\big)$:

$$(X,\mathfrak{A}) \xrightarrow{\;f\;} (\mathbb{R},\mathfrak{Bor}_{\mathbb{R}})$$

$$\big(X,\mathfrak{G}\big) \xrightarrow{\;e\;} (\mathbb{R},\mathfrak{Bor}_{\mathbb{R}}) \quad \text{meaning} \quad E\big(\overline{f}^{\mu}\,|\,g\big) = \overline{r\circ g}^{\mu_{|\mathfrak{G}}}$$

$$\big(X,\mathfrak{G}\big) \xrightarrow{\;g\;} (Y,\mathfrak{A}^{(g)}) \quad\text{with}\quad (Y,\mathfrak{A}^{(g)}) \xrightarrow{\;r\;} (\mathbb{R},\mathfrak{Bor}_{\mathbb{R}})$$

Call the induced function r the **regression function**.

Discussion 7.6. In a special case of Definition 7.5, if the function $g : X \to Y$ is known to be $(\mathfrak{A}, \mathfrak{B})$-measurable, then $\mathfrak{A}^{(g)} = \mathfrak{B}$ and

$$\mathfrak{G} := g^{\leftarrow}(\mathfrak{B}) \cap \mathfrak{A} = g^{\leftarrow}(\mathfrak{B}).$$

Which means in this case that the conditional expectation $E\big(\overline{f}^{\mu} \,|\, g\big)$ of $\overline{f}^{\mu}$ given the function g denotes the same class as the conditional expectation $E_{\mathfrak{G}}\, \overline{f}^{\mu}$ of $\overline{f}^{\mu}$ with respect to the σ-algebra $\mathfrak{G}$.

7.1 Fundamental Properties

Definition 7.7. Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The following defines a contractive operator:

$$E_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}).$$

By Proposition 4.14, there is another contractive operator:

$$\mathfrak{I} : L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \rightarrowtail L_1(X, \mathfrak{A}, \mu) : \overline{f}^{\mu_{|\mathfrak{B}}} \mapsto \overline{f}^{\mu}.$$

Define the **conditional expectation operator**

$$\mathscr{E}_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1(X, \mathfrak{A}, \mu)$$

to be their composition $\mathfrak{I} \circ E_{\mathfrak{B}}$:

$$\begin{array}{ccc} L_1(X, \mathfrak{A}, \mu) & \xrightarrow{\mathscr{E}_{\mathfrak{B}}} & L_1(X, \mathfrak{A}, \mu) \\ & \searrow^{E_{\mathfrak{B}}} & \uparrow \mathfrak{I} \\ & & L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \end{array}$$

Proposition 7.8. *The map $\mathfrak{I}$ preserves the integral, meaning if $\overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, then*

$$\int_B \mathscr{E}_{\mathfrak{B}}\, \overline{g}^{\mu}\, d\mu = \int_B E_{\mathfrak{B}}\, \overline{g}^{\mu}\, d\mu_{|\mathfrak{B}} \quad \textit{for all } B \in \mathfrak{B}.$$

Remark 7.9. For each class $\overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, the class $\mathscr{E}_{\mathfrak{B}}\, \overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$ has a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative, and to see this, say the function $e \in \mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ is the $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative of the class $E_{\mathfrak{B}}\, \overline{g}^{\mu}$, as in Definition 7.3. This means $E_{\mathfrak{B}}\, \overline{g}^{\mu} = \overline{e}^{\mu_{|\mathfrak{B}}}$. Then the equalities

$$\mathscr{E}_{\mathfrak{B}}\, \overline{g}^{\mu} = \mathfrak{I}\big(E_{\mathfrak{B}}\, \overline{g}^{\mu}\big) = \mathfrak{I}\big(\overline{e}^{\mu_{|\mathfrak{B}}}\big) = \overline{e}^{\mu}$$

show that $\mathscr{E}_{\mathfrak{B}}\, \overline{g}^{\mu}$ also has the function e, as included in $\mathscr{L}_1(X, \mathfrak{A}, \mu)$, as a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative. You may recall the inclusion $\mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \subseteq \mathscr{L}_1(X, \mathfrak{A}, \mu)$ of Discussion 4.16.

Proposition 7.10. *Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The conditional expectation operator*

$$\mathscr{E}_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1(X, \mathfrak{A}, \mu) : \overline{f}^{\mu} \mapsto \mathscr{E}_{\mathfrak{B}}\, \overline{f}^{\mu}$$

is positive, linear, and contractive.

> The correct and complete proof of the functional equation $\mathsf{E}\, \overline{f}^{\mu} = \mathsf{E}\, \big(\mathscr{E}_{\mathfrak{B}}\, \overline{f}^{\mu}\big)$ in the following proposition took some effort, as tracing back through at least Proposition 4.14, shows. Let my early and admittedly naive attempts to try to simply throw a monotone convergence theorem at this proposition at least bring a smile to your face. Of course there was more to it.

Proposition 7.11. *If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, then*

$$\mathsf{E}\, \overline{f}^{\mu} = \mathsf{E}\, \big(\mathscr{E}_{\mathfrak{B}}\, \overline{f}^{\mu}\big)$$

for all classes $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$.

Proof. Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. Let $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. By definition, $\mathsf{E}\,\overline{f}^{\mu} = \int_X f\,d\mu$. Also by definition, $\mathsf{E}\left(\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\right) = \bar{\int}_X \mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\,d\mu$. We will show that these two integrals are equal.

By definition of conditional mathematical expectation, we know that

$$\int_B f\,d\mu = \bar{\int}_B E_{\mathfrak{B}}\,\overline{f}^{\mu}\,d\mu_{|\mathfrak{B}}$$

for all $B \in \mathfrak{B}$, and by taking $B = X$, it follows that

$$\int_X f\,d\mu = \bar{\int}_X E_{\mathfrak{B}}\,\overline{f}^{\mu}\,d\mu_{|\mathfrak{B}}.$$

By Proposition 7.8,

$$\bar{\int}_X E_{\mathfrak{B}}\,\overline{f}^{\mu}\,d\mu_{|\mathfrak{B}} = \bar{\int}_X \mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\,d\mu.$$

It follows by transitivity of equality that

$$\int_X f\,d\mu = \bar{\int}_X \mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\,d\mu,$$

which means $\mathsf{E}\,\overline{f}^{\mu} = \mathsf{E}\left(\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\right)$, as required. ■

Proposition 7.12. *If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, then*

$$E_{\mathfrak{B}}\,\mathfrak{I}(\overline{f}^{\mu_{|\mathfrak{B}}}) = \overline{f}^{\mu_{|\mathfrak{B}}}$$

for all classes $\overline{f}^{\mu_{|\mathfrak{B}}} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.

Proof. From the definition of conditional mathematical expectation: if $\mathfrak{I}(\overline{f}^{\mu_{|\mathfrak{B}}})$ in $L_1(X, \mathfrak{A}, \mu)$, there is a unique class $E_{\mathfrak{B}}\,\mathfrak{I}(\overline{f}^{\mu_{|\mathfrak{B}}})$ in $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ such that

$$\bar{\int}_B \mathfrak{I}(\overline{f}^{\mu_{|\mathfrak{B}}})\,d\mu = \bar{\int}_B E_{\mathfrak{B}}\,\mathfrak{I}(\overline{f}^{\mu_{|\mathfrak{B}}})\,d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B}.$$

It follows by Proposition 4.14 that

$$\int_B \bar{\mathfrak{I}}(\overline{f}^{\mu_{|\mathfrak{B}}})\, d\mu = \int_B \overline{f}^{\mu_{|\mathfrak{B}}}\, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B}.$$

By uniqueness, $E_{\mathfrak{B}}\,\bar{\mathfrak{I}}(\overline{f}^{\mu_{|\mathfrak{B}}}) = \overline{f}^{\mu_{|\mathfrak{B}}}$, completing the proof. ■

Remark 7.13. Proposition 7.12 says that the following diagram commutes:

$$\begin{array}{ccc} L_1(X,\mathfrak{A},\mu) & & \\ \Big\uparrow \mathfrak{I} & \searrow^{E_{\mathfrak{B}}} & \\ L_1(X,\mathfrak{B},\mu_{|\mathfrak{B}}) & \xrightarrow{\text{identity}} & L_1(X,\mathfrak{B},\mu_{|\mathfrak{B}}) \end{array}$$

which is part of a larger commutative diagram:

$$\begin{array}{ccccc} L_1(X,\mathfrak{A},\mu) & & \xrightarrow{\mathscr{E}_{\mathfrak{B}}} & & L_1(X,\mathfrak{A},\mu) \\ \Big\uparrow \mathfrak{I} & \searrow^{E_{\mathfrak{B}}} & & & \Big\uparrow \mathfrak{I} \\ L_1(X,\mathfrak{B},\mu_{|\mathfrak{B}}) & & \xrightarrow{\text{identity}} & & L_1(X,\mathfrak{B},\mu_{|\mathfrak{B}}) \end{array}$$

7.2 Averaging Properties

Definition 7.14. Let $(X,\mathfrak{S},\mu)$ be a measure space. A μ^*-measurable set A with $\mu^*(A) > 0$ is called an **atom** of $(X,\mathfrak{S},\mu)$ if for every μ^*-measurable subset B of A either $\mu^*(B) = 0$, or $\mu^*(A \cap B^{\mathsf{c}}) = 0$.

Proposition 7.15. *Let $(X,\mathfrak{S},\mu)$ be a finite measure space. A μ^*-measurable set A with $\mu^*(A) > 0$ is an atom of $(X,\mathfrak{S},\mu)$ if for every μ^*-measurable subset B of A either $\mu^*(B) = \mu^*(A)$, or $\mu^*(A \cap B^{\mathsf{c}}) = \mu^*(A)$.*

Remark 7.16. If the measure space $(X, \mathfrak{S}, \mu)$ is not finite, then the condition that either $\mu^*(B) = \mu^*(A)$, or $\mu^*(A \cap B^{\mathsf{C}}) = \mu^*(A)$ is not strong enough to imply the condition that either $\mu^*(B) = 0$, or $\mu^*(A \cap B^{\mathsf{C}}) = 0$.

Proposition 7.17. *Let $(X, \mathfrak{A}, \mu)$ be a probability space. Let $\mathfrak{B}$ denote the σ-subalgebra of $\mathfrak{A}$ generated by a countable partition $\{B_i\}$ of X, where each set $B_i \in \mathfrak{A}$. Then every set B_i with $\mu(B_i) > 0$ is an atom of the probability space $(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.*

Proof. Let us suppose that the set B_1 has $\mu(B_1) > 0$, and let us show that this implies B_1 is an atom of $(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. We first need to show both that B_1 is a $\mu_{|\mathfrak{B}}^*$-measurable set, and that $\mu_{|\mathfrak{B}}^*(B_1) > 0$. We then need to show (by Proposition 7.15) that if A is a $\mu_{|\mathfrak{B}}^*$-measurable subset of B_1, then either $\mu_{|\mathfrak{B}}^*(A) = \mu_{|\mathfrak{B}}^*(B_1)$, or $\mu_{|\mathfrak{B}}^*(B_1 \cap A^{\mathsf{C}}) = \mu_{|\mathfrak{B}}^*(B_1)$.

First, by Proposition 2.16, we know that $\mathfrak{B} \subseteq \mu_{|\mathfrak{B}}^*$-$\mathfrak{Meas}$, and since $B_1 \in \mathfrak{B}$, it follows that B_1 is a $\mu_{|\mathfrak{B}}^*$-measurable subset of X. Also, by Proposition 3.14 and Proposition 2.11, we know that $\mu(B_1) = \mu_{|\mathfrak{B}}(B_1) = \mu_{|\mathfrak{B}}^*(B_1)$, and so $P(B_1) > 0$ means that $\mu_{|\mathfrak{B}}^*(B_1) > 0$.

Next, let A be a $\mu_{|\mathfrak{B}}^*$-measurable subset of B_1. We need to see that either $\mu_{|\mathfrak{B}}^*(A) = \mu_{|\mathfrak{B}}^*(B_1)$, or $\mu_{|\mathfrak{B}}^*(B_1 \cap A^{\mathsf{C}}) = \mu_{|\mathfrak{B}}^*(B_1)$. Should A be empty, then $B_1 \cap A^{\mathsf{C}} = B_1$, and so $\mu_{|\mathfrak{B}}^*(B_1 \cap A^{\mathsf{C}}) = \mu_{|\mathfrak{B}}^*(B_1)$, rather vacuously. Suppose then that A is non-empty, and let us see that this implies $\mu_{|\mathfrak{B}}^*(A) = \mu_{|\mathfrak{B}}^*(B_1)$. Let us determine $\mu_{|\mathfrak{B}}^*(A)$ by tending to the definition:

$$\mu_{|\mathfrak{B}}^*(A) = \inf\left\{\sum_{n=1}^{\infty} \mu_{|\mathfrak{B}}(A_n) : \{A_n\} \text{ is a sequence of } \mathfrak{B} \text{ with } A \subseteq \bigcup_{n=1}^{\infty} A_n\right\}.$$

If $\{A_n\}$ is a sequence of $\mathfrak{B}$ with $A \subseteq \bigcup_{n=1}^{\infty} A_n$, then because $\mathfrak{B}$ is generated by a countable collection of disjoint sets B_i, we may assume that each of the A_n is either empty, or equal to one of the B_i. Furthermore, since A is non-empty, at least one of the A_n must equal B_1 in order that $A \subseteq \bigcup_{n=1}^{\infty} A_n$. This means that for any sequence, $\{C_n\}$ say, of $\mathfrak{B}$ where exactly one of the C_n is equal to B_1 and the rest of the C_n are empty, we must have

$$\sum_{n=1}^{\infty} \mu_{|\mathfrak{B}}(C_n) \leq \sum_{n=1}^{\infty} \mu_{|\mathfrak{B}}(A_n)$$

if $\{A_n\}$ is an arbitrary sequence of $\mathfrak{B}$ with $A \subseteq \bigcup_{n=1}^{\infty} A_n$. And $\sum_{n=1}^{\infty} \mu_{|\mathfrak{B}}(C_n) = \mu_{|\mathfrak{B}}(B_1) = \mu_{|\mathfrak{B}}^*(B_1)$. So $\mu_{|\mathfrak{B}}^*(B_1)$ is a lower bound for the set

$$\left\{ \sum_{n=1}^{\infty} \mu_{|\mathfrak{B}}(A_n) : \{A_n\} \text{ is a sequence of } \mathfrak{B} \text{ with } A \subseteq \bigcup_{n=1}^{\infty} A_n \right\}.$$

But $\mu_{|\mathfrak{B}}^*(A)$ is the greatest such lower bound, so $\mu_{|\mathfrak{B}}^*(A) \geq \mu_{|\mathfrak{B}}^*(B_1)$. On the other hand, $A \subseteq B_1$, so $\mu_{|\mathfrak{B}}^*(A) \leq \mu_{|\mathfrak{B}}^*(B_1)$. It follows then that $\mu_{|\mathfrak{B}}^*(A) = \mu_{|\mathfrak{B}}^*(B_1)$, completing the proof. ■

Lemma 7.18. *If $f : X \to \mathbb{R}$ is an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple function, then f is constant μ^*-almost everywhere on each atom of the probability space $(X, \mathfrak{A}, \mu)$.*

Proof. Let $f : X \to \mathbb{R}$ be an $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-simple function, meaning that f is a $(\mathfrak{A}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function having finite range. Suppose the range of f to be $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}$. If $n = 1$, then f is certainly constant since $f = a_1$ on X, so suppose $n > 1$. Define subsets $A_i \subseteq X$ to be the preimages $f^{\leftarrow}(a_i)$. Let A be an atom of $(X, \mathfrak{A}, \mu)$. The subsets $\{A_i\}$ partition X, and so the atom A can be written as

the disjoint union $\bigcup_{i=1}^{n}(A_i \cap A)$, where each $A_i \cap A$ is a μ^*-measurable subset of A. Since μ^* is σ-additive on the μ^*-measurable subset of A (by Proposition 2.14), it follows that $\mu^*(A) = \sum_{i=1}^{n} \mu^*(A_i \cap A)$. Because A is an atom, it follows (by Proposition 7.15) that for each A_i, either $\mu^*(A_i \cap A) = \mu^*(A)$, or $\mu^*\big(\bigcup_{j\neq i}(A_j \cap A)\big) = \mu^*(A)$. Let us argue that $\mu^*(A) = \mu^*(A_i \cap A)$ for exactly one of the A_i, for it will then follow that $f = a_i$ μ^*-almost everywhere on A.

Should $\mu^*(A) = \mu^*(A_i \cap A)$ for at least two of the A_i; say for A_1 and A_2, we would then have

$$\mu^*(A) = \sum_{i=1}^{n} \mu^*(A_i \cap A) \geq \mu^*(A_1 \cap A) + \mu^*(A_2 \cap A) = 2 \cdot \mu^*(A),$$

which could only happen were $\mu^*(A) = 0$. But $\mu^*(A) > 0$, by assumption.

Should $\mu^*(A) \neq \mu^*(A_i \cap A)$ for any of the A_i, then $\mu^*\big(\bigcup_{j\neq i}(A_j \cap A)\big) = \mu^*(A)$ for each A_i, and so

$$\begin{aligned}\mu^*(A) &= \mu^*\Big((A_i \cap A) \cup \big(\textstyle\bigcup_{j\neq i}(A_j \cap A)\big)\Big) \\ &= \mu^*(A_i \cap A) + \mu^*\big(\textstyle\bigcup_{j\neq i}(A_j \cap A)\big) \\ &= \mu^*(A_i \cap A) + \mu^*(A),\end{aligned}$$

meaning $\mu^*(A_i \cap A) = 0$ for each i. This would imply that $\mu^*(A) = 0$ because $\mu^*(A) = \sum_{i=1}^{n} \mu^*(A_i \cap A)$. But $\mu^*(A) > 0$ by assumption.

Conclude that $\mu^*(A) = \mu^*(A_i \cap A)$ for exactly one of the A_i, so $\mu^*(A_j \cap A) = 0$ for $j \neq i$, and consequently $f = a_i$ μ^*-almost everywhere on A, completing the proof. ■

Proposition 7.19. *If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, and if $f : X \to \mathbb{R}$ is a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function, then f is constant $\mu_{|\mathfrak{B}}^*$-almost everywhere on each atom of the probability space $(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.*

Proof. Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$, and let $f : X \to \mathbb{R}$ be $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. Also, let B be an atom of $(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. By Proposition 3.21, there is a sequence $\{f_n\}$ of $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-simple functions on X such that $\lim f_n(x) = f(x)$ for all $x \in X$. Because the f_n converge to f on X, and because each f_n is constant $\mu_{|\mathfrak{B}}^*$-almost everywhere on B (by Lemma 7.18), it follows that f is constant $\mu_{|\mathfrak{B}}^*$-almost everywhere on B, as required. ■

Discussion 7.20. If $e \in \mathscr{L}_1(X, \mathfrak{A}, \mu)$, and if $B \in \mathfrak{A}$ so that $\chi_B \in \mathscr{L}_1(X, \mathfrak{A}, \mu)$, and if the product of functions $e\chi_B$ is in $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ so that the class $\overline{e\chi_B}^{\mu}$ is in $L_1(X, \mathfrak{A}, \mu)$, then the product $\overline{e}^{\mu}\overline{\chi_B}^{\mu}$ of classes is defined to be the class $\overline{e\chi_B}^{\mu}$ in $L_1(X, \mathfrak{A}, \mu)$. Should the function e be any representative of the class $\mathscr{E}_{\mathfrak{B}}\,\overline{g}^{\mu}$, then $\overline{e}^{\mu} = \mathscr{E}_{\mathfrak{B}}\,\overline{g}^{\mu}$ and so $\overline{e}^{\mu}\overline{\chi_B}^{\mu} = (\mathscr{E}_{\mathfrak{B}}\,\overline{g}^{\mu})\overline{\chi_B}^{\mu}$ in $L_1(X, \mathfrak{A}, \mu)$.

Proposition 7.21. *Let $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, and if B is an atom of the probability space $(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, then*

$$\left(\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\right)\overline{\chi_B}^{\mu} = \left(\frac{1}{\mu(B)}\int_B f\,d\mu\right)\overline{\chi_B}^{\mu} \quad \textit{in } L_1(X, \mathfrak{A}, \mu).$$

Proof. The idea is to show that

$$\left(E_{\mathfrak{B}}\,\overline{f}^{\mu}\right)\overline{\chi_B}^{\mu_{|\mathfrak{B}}} = \left(\frac{1}{\mu(B)}\int_B f\,d\mu\right)\overline{\chi_B}^{\mu_{|\mathfrak{B}}} \quad \text{in } L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}),$$

and then apply the map $\mathfrak{I}$ of Proposition 4.14.

Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$, and let B be an atom of $(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. We may suppose the function $e \in \mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ to be the $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable representative of the class $E_{\mathfrak{B}}\,\overline{f}^{\mu} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. It follows by Proposition 7.19 that the function e is constant $\mu_{|\mathfrak{B}}^*$-almost everywhere on B. Suppose this constant equals a, meaning

that the value of e is equal to a $\mu_{|\mathfrak{B}}^*$-almost everywhere on B. Constants are easy to integrate over sets:

$$\int_B e\, d\mu_{|\mathfrak{B}} = a \cdot \mu_{|\mathfrak{B}}(B) = a \cdot \mu(B).$$

On the other hand, by the definition of conditional expectation,

$$\int_B e\, d\mu_{|\mathfrak{B}} = \int_B E_{\mathfrak{B}}\, \overline{f}^{\mu}\, d\mu_{|\mathfrak{B}} = \int_B f\, d\mu.$$

This says that

$$a \cdot \mu(B) = \int_B f\, d\mu.$$

By hypothesis, the set B is an atom, and so $\mu(B) > 0$. This means we can divide by $\mu(B)$. Therefore

$$a = \frac{1}{\mu(B)} \int_B f\, d\mu.$$

Because e is equal to a $\mu_{|\mathfrak{B}}^*$-almost everywhere on B, it follows that

$$e = \frac{1}{\mu(B)} \int_B f\, d\mu \quad \mu_{|\mathfrak{B}}^*\text{-almost everywhere on } B,$$

and so

$$e\chi_B = \left(\frac{1}{\mu(B)} \int_B f\, d\mu\right) \chi_B \quad \mu_{|\mathfrak{B}}^*\text{-almost everywhere on } X.$$

By taking equivalence classes in $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, we get

$$\overline{e\chi_B}^{\mu_{|\mathfrak{B}}} = \left(\frac{1}{\mu(B)} \int_B f\, d\mu\right) \overline{\chi_B}^{\mu_{|\mathfrak{B}}} \quad \text{in } L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}).$$

Since $\overline{e\chi_B}^{\mu_{|\mathfrak{B}}} = \left(E_{\mathfrak{B}}\,\overline{f}^{\mu}\right)\overline{\chi_B}^{\mu_{|\mathfrak{B}}}$ in $L_1(X,\mathfrak{B},\mu_{|\mathfrak{B}})$, by substitution

$$\left(E_{\mathfrak{B}}\,\overline{f}^{\mu}\right)\overline{\chi_B}^{\mu_{|\mathfrak{B}}} = \left(\frac{1}{\mu(B)}\int_B f\,d\mu\right)\overline{\chi_B}^{\mu_{|\mathfrak{B}}} \quad \text{in } L_1(X,\mathfrak{B},\mu_{|\mathfrak{B}}).$$

Applying the map $\mathfrak{I}$, we get

$$\left(\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu}\right)\overline{\chi_B}^{\mu} = \left(\frac{1}{\mu(B)}\int_B f\,d\mu\right)\overline{\chi_B}^{\mu} \quad \text{in } L_1(X,\mathfrak{A},\mu),$$

completing the proof. ■

Proposition 7.22. *Let $(X,\mathfrak{A},\mu)$ be a probability space, and suppose that $\mathfrak{A}$ has a σ-subalgebra $\mathfrak{B}$ generated by a countable partition $\{B_i\}$ of X, where each $B_i \in \mathfrak{A}$ and $\mu(B_i) > 0$. If $\overline{f}^{\mu} \in L_1(X,\mathfrak{A},\mu)$, then*

$$\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu} = \sum_{i=1}^{\infty}\left(\frac{1}{\mu(B_i)}\int_{B_i} f\,d\mu\right)\overline{\chi_{B_i}}^{\mu} \quad \textit{in } L_1(X,\mathfrak{A},\mu).$$

Proof. This simply combines Propositions 7.17 and 7.21. ■

Notation 7.23. Let 1_X denote the constant function $X \to \mathbb{R} : x \mapsto 1$.

Proposition 7.24. *Let $\overline{f}^{\mu} \in L_1(X,\mathfrak{A},\mu)$. If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, and if the σ-subalgebras $\sigma(f)$ and $\mathfrak{B}$ of the σ-algebra $\mathfrak{A}$ are independent with respect to μ^*, then*

$$\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu} = \left(\mathsf{E}\,\overline{f}^{\mu}\right)\left(\overline{1_X}^{\mu}\right) \quad \textit{in } L_1(X,\mathfrak{A},\mu).$$

Proof. In order to show that the equality $\mathscr{E}_{\mathfrak{B}}\,\overline{f}^{\mu} = \left(\mathsf{E}\,\overline{f}^{\mu}\right)\left(\overline{1_X}^{\mu}\right)$ holds in $L_1(X,\mathfrak{A},\mu)$, it's sufficient to show that $E_{\mathfrak{B}}\,\overline{f}^{\mu} = \left(\mathsf{E}\,\overline{f}^{\mu}\right)\left(\overline{1_X}^{\mu_{|\mathfrak{B}}}\right)$

holds in $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, and in turn, to show this, it's sufficient to show

$$\int_B E_{\mathfrak{B}}\, \overline{f}^{\mu}\, d\mu_{|\mathfrak{B}} = \int_B \big(\mathsf{E}\, \overline{f}^{\mu}\big)\big(\overline{1_X}^{\mu_{|\mathfrak{B}}}\big)\, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B}.$$

The independence of the σ-subalgebras $\sigma(f)$ and $\mathfrak{B}$ with respect to μ^* implies that the classes $\overline{f}^{\mu}$ and $\overline{\chi_B}^{\mu}$ are independent with respect to μ^* for all $B \in \mathfrak{B}$. And, as an application of Proposition 4.14,

$$\mathsf{E}\, \overline{\chi_B}^{\mu} = \int_X \chi_B\, d\mu = \int_X \chi_B\, d\mu_{|B} = \int_B \overline{1_X}^{\mu_{|\mathfrak{B}}}\, d\mu_{|\mathfrak{B}}.$$

Consequently,

$$\begin{aligned}
\int_B E_{\mathfrak{B}}\, \overline{f}^{\mu}\, d\mu_{|\mathfrak{B}} &= \int_{\mathfrak{B}} f\, d\mu && \text{by definition} \\
&= \int_X f\chi_B\, d\mu \\
&= \mathsf{E}\,\big(\overline{f\chi_B}^{\mu}\big) \\
&= \mathsf{E}\,\big(\overline{f}^{\mu}\overline{\chi_B}^{\mu}\big) \\
&= \big(\mathsf{E}\, \overline{f}^{\mu}\big)\big(\mathsf{E}\, \overline{\chi_B}^{\mu}\big) && \text{by Prop. 6.12} \\
&= \big(\mathsf{E}\, \overline{f}^{\mu}\big) \int_B \overline{1_X}^{\mu_{|\mathfrak{B}}}\, d\mu_{|\mathfrak{B}} && \text{by substitution} \\
&= \int_B \big(\mathsf{E}\, \overline{f}^{\mu}\big)\big(\overline{1_X}^{\mu_{|\mathfrak{B}}}\big)\, d\mu_{|\mathfrak{B}},
\end{aligned}$$

as required. ■

Discussion 7.25. On any probability space $(X, \mathfrak{A}, \mu)$, the σ-algebra generated by the constant function 1_X is $\{X, \varnothing\}$, which is independent of every σ-subalgebra of $\mathfrak{A}$. It follows by Proposition 7.24

that $\mathscr{E}_{\sigma(1_X)}\,\overline{f}^{\mu} = \big(\mathsf{E}\,\overline{f}^{\mu}\big)\big(\overline{1_X}^{\mu}\big)$ for every $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. This means that any equation involving conditional mathematical expectation which holds for all σ-subalgebras of $\mathfrak{A}$ must hold for expected values as well. You may argue that this is logically backwards since the conditional mathematical expectation is a generalization of the expected value.

Proposition 7.26. *If classes $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ in $L_1(X, \mathfrak{A}, \mu)$ are independent, then*

$$\mathscr{E}_{\sigma(g)}\,\overline{f}^{\mu} = \big(\mathsf{E}\,\overline{f}^{\mu}\big)\big(\overline{1_X}^{\mu}\big)$$

and

$$\mathscr{E}_{\sigma(f)}\,\overline{g}^{\mu} = \big(\mathsf{E}\,\overline{g}^{\mu}\big)\big(\overline{1_X}^{\mu}\big).$$

Proof. These are just specific instances of Proposition 7.24. ■

Remark 7.27. A commutative diagram is reproduced here to keep track of what's going on in the following proposition:

$$\begin{array}{ccc} L_1(X, \mathfrak{A}, \mu) & \xrightarrow{\mathscr{E}_{\mathfrak{B}}} & L_1(X, \mathfrak{A}, \mu) \\ \Big\uparrow \mathfrak{I} & \searrow^{E_{\mathfrak{B}}} & \Big\uparrow \mathfrak{I} \\ L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) & \xrightarrow{\text{identity}} & L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \end{array}$$

Proposition 7.28. *Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$, and let the class $\overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. If $f : X \to \mathbb{R}$ is a $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable function and f is bounded μ^*-almost everywhere on X, and if the class $\overline{f}^{\mu}\overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, then*

$$E_{\mathfrak{B}}\big(\overline{f}^{\mu}\overline{g}^{\mu}\big) = \overline{f}^{\mu_{|\mathfrak{B}}}\, E_{\mathfrak{B}}\,\overline{g}^{\mu},$$

and consequently

$$\mathscr{E}_{\mathfrak{B}}\big(\overline{f}^{\mu}\overline{g}^{\mu}\big) = \overline{f}^{\mu}\,\mathscr{E}_{\mathfrak{B}}\,\overline{g}^{\mu}.$$

Proof. Let the function $e \in \mathscr{L}_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ represent the class $E_B\,\overline{g}^{\mu} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, so that $\overline{e}^{\mu_{|\mathfrak{B}}} = E_B\,\overline{g}^{\mu}$, and let the set $C \in \mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas} \subseteq \mu^*\text{-}\mathfrak{Meas}$. We will show that $\int_B \chi_C g\,d\mu = \overline{\int}_B \overline{\chi_C e}^{\mu_{|\mathfrak{B}}}\,d\mu_{|\mathfrak{B}}$ for all $B \in \mathfrak{B}$, which implies, by the uniqueness of conditional mathematical expectation of $\overline{\chi_C g}^{\mu}$ with respect to $\mathfrak{B}$, that $E_{\mathfrak{B}}\,\overline{\chi_C g}^{\mu} = \overline{\chi_C e}^{\mu_{|\mathfrak{B}}}$. See that

$$\begin{aligned}
\int_B \chi_C g\,d\mu &= \int_{B\cap C} g\,d\mu \\
&= \overline{\int}_{B\cap C} E_{\mathfrak{B}}\,\overline{g}^{\mu}\,d\mu_{|\mathfrak{B}} \\
&= \int_{B\cap C} e\,d\mu_{|\mathfrak{B}} \\
&= \int_B \chi_C e\,d\mu_{|\mathfrak{B}} \\
&= \overline{\int}_B \overline{\chi_C e}^{\mu_{|\mathfrak{B}}}\,d\mu_{|\mathfrak{B}}.
\end{aligned}$$

This shows that

$$E_{\mathfrak{B}}\,\overline{\chi_C g}^{\mu} = \overline{\chi_C e}^{\mu_{|\mathfrak{B}}},$$

and since the equivalence class $\overline{\chi_C e}^{\mu_{|\mathfrak{B}}}$ can be expressed as

$$\begin{aligned}
\overline{\chi_C e}^{\mu_{|\mathfrak{B}}} &= \overline{\chi_C}^{\mu_{|\mathfrak{B}}}\overline{e}^{\mu_{|\mathfrak{B}}} \\
&= \overline{\chi_C}^{\mu_{|\mathfrak{B}}}\,E_{\mathfrak{B}}\,\overline{g}^{\mu},
\end{aligned}$$

it follows that

$$E_{\mathfrak{B}}\big(\overline{\chi_C}^{\mu}\overline{g}^{\mu}\big) = \overline{\chi_C}^{\mu_{|\mathfrak{B}}}\,E_{\mathfrak{B}}\,\overline{g}^{\mu}.$$

This last equality holds for all $C \in \mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas}$, and so by linearity for all equivalence classes of $(\mu_{|\mathfrak{B}}^*\text{-}\mathfrak{Meas}, \mathfrak{Bor}_{\mathbb{R}})$-measurable step functions.

The result follows by a continuity argument, which fortunately works well in a normed space. Not so well in a semi-normed space. ■

Proposition 7.29. *Let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$, and let the class $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. If the function f is $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, then*

$$\overline{f}^{\mu} = \mathscr{E}_{\mathfrak{B}}\, \overline{f}^{\mu}.$$

Proof. Take $\overline{g}^{\mu} = \overline{1_X}^{\mu}$, and apply the previous proposition. ■

8 Conditional Probability and Markov Kernels

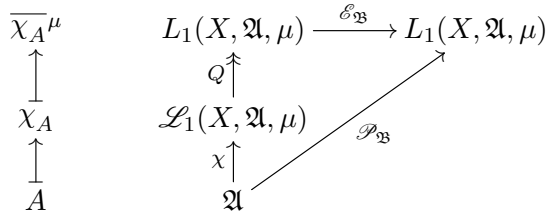
SOME authors confuse Markov kernels and conditional distributions. We will not. However, as shown in Proposition 8.21 below, a Markov kernel satisfying an extra condition corresponds to a regular conditional distribution.

8.1 Conditional Probability

Definition 8.1 (Conditional probability). Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The set function

$$\mathscr{P}_{\mathfrak{B}} : \mathfrak{A} \to L_1(X, \mathfrak{A}, \mu) : A \mapsto \mathscr{E}_{\mathfrak{B}}\, \overline{\chi_A}^{\mu}$$

is the **conditional probability** on $\mathfrak{A}$ with respect to the σ-subalgebra $\mathfrak{B}$, or the *conditional probability on* $\mathfrak{A}$ *given* $\mathfrak{B}$, the composition being illustrated in the following diagram:



Remark 8.2. Should the σ-subalgebra $\mathfrak{B}$ be generated by a countable partition $\{B_i\}$ of X, where each $B_i \in \mathfrak{A}$ and $\mu(B_i) > 0$, then using Proposition 7.22, for each $A \in \mathfrak{A}$,

$$\begin{aligned}\mathscr{P}_{\mathfrak{B}}(A) &= \mathscr{E}_{\mathfrak{B}}\,\overline{\chi_A}^{\mu} \\ &= \sum_{i=1}^{\infty}\left(\frac{1}{\mu(B_i)}\int_{B_i}\chi_A\,d\mu\right)\overline{\chi_{B_i}}^{\mu} \\ &= \sum_{i=1}^{\infty}\left(\frac{\mu(A\cap B_i)}{\mu(B_i)}\right)\overline{\chi_{B_i}}^{\mu}.\end{aligned}$$

This reminds us of the symbols in the very elementary definition of conditional probability:

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

Proposition 8.3. *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The conditional probability $\mathscr{P}_{\mathfrak{B}} : \mathfrak{A} \to L_1(X, \mathfrak{A}, \mu)$ has the following measure-like properties:*

1. $\mathscr{P}_{\mathfrak{B}}(\varnothing) = \overline{0_X}^{\mu}$.
2. *$\mathscr{P}_{\mathfrak{B}}$ is σ-additive.*
3. $\mathscr{P}_{\mathfrak{B}}(X) = \overline{1_X}^{\mu}$.

8.2 Markov Kernels

Notation 8.4. Let X be a space and $\mathfrak{B}$ be a σ-algebra. For each set B in the σ-algebra $\mathfrak{B}$, define the function

$$\mathsf{fix}^B : X \to X \times \mathfrak{B} : x \mapsto (x, B),$$

and for each element x in the space X, define the function

$$\mathsf{fix}_x : \mathfrak{B} \to X \times \mathfrak{B} : B \mapsto (x, B).$$

Definition 8.5. Given measurable spaces $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$, a **Markov kernel** is a function

$$k : X \times \mathfrak{B} \to [0, 1]$$

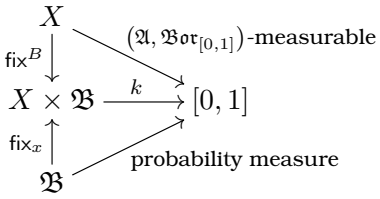
such that for each set $B \in \mathfrak{B}$, the map

$$k \circ \mathsf{fix}^B : X \to X \times \mathfrak{B} \to [0, 1] : x \mapsto (x, B) \mapsto k(x, B)$$

is $\big(\mathfrak{A}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, and for each element $x \in X$, the map

$$k \circ \mathsf{fix}_x : \mathfrak{B} \to X \times \mathfrak{B} \to [0, 1] : B \mapsto (x, B) \mapsto k(x, B)$$

is a probability measure on $\mathfrak{B}$. It means given an $x \in X$ and a $B \in \mathfrak{B}$ there is a commutative diagram, only one of which will be illustrated without quantification:



The given measurable spaces $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ need not be distinct, which may be apparent by the domain of k. A subscript indicating a domain, such as fix^B_X, could be used in case more than one kernel is involved. The map $k \circ \mathsf{fix}^B$ may be denoted with something like k^B, and likewise the map $k \circ \mathsf{fix}_x$ may be denoted with something like k_x.

A Markov kernel given measurable spaces $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ is also called a **transition** from $(X, \mathfrak{A})$ to $(Y, \mathfrak{B})$, possibly using the notation $k : (X, \mathfrak{A}) \prec (Y, \mathfrak{B})$ in place of specifying $k : X \times \mathfrak{B} \to [0, 1]$.

Example 8.6. Every measurable function induces a Markov kernel, and to see this, let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces, and let $f : X \to Y$ be an $(\mathfrak{A}, \mathfrak{B})$-measurable function. Define a function $k : X \times \mathfrak{B} \to [0, 1]$ by

$$(x, B) \mapsto \begin{cases} 1 & \text{if } f(x) \in B \\ 0 & \text{if } f(x) \notin B. \end{cases}$$

Two other ways to view the value $k(x, B)$ would be as $k(x, B) = \chi_B\big(f(x)\big)$, and as $k(x, B) = \delta_{f(x)}(B)$ with $\delta_{f(x)}$ denoting Dirac measure concentrated at $f(x)$. Then for each set $B \in \mathfrak{B}$, the map

$$k \circ \mathsf{fix}^B : X \to [0, 1] : x \mapsto k(x, B) = \chi_B\big(f(x)\big)$$

can be seen as a composition of the $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable function χ_B with the $(\mathfrak{A}, \mathfrak{B})$-measurable function f, and so the composition is $\big(\mathfrak{A}, \mathfrak{Bor}_{[0,1]}\big)$-measurable. And for each element $x \in X$, the map

$$k \circ \mathsf{fix}_x : \mathfrak{B} \to [0, 1] : B \mapsto k(x, B) = \delta_{f(x)}(B)$$

visibly defines a measure on $\mathfrak{B}$, in particular Dirac measure concentrated at $f(x)$. And since $f(x) \in Y$ implies that the Dirac measure of the whole space Y is 1, the map $k \circ \mathsf{fix}_x$ further defines a probability measure on $\mathfrak{B}$. This shows that k is a Markov kernel by the very definition of Markov kernel.

Remark 8.7. The following proposition could be stated using a Markov kernel.

Proposition 8.8. *Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces. Let μ be a probability measure on $\mathfrak{A}$. If $\{k^B : B \in \mathfrak{B}\}$ is a collection of $\big(\mathfrak{A}, \mathfrak{Bor}_{[0,1]}\big)$-measurable functions, and $\{k_x : x \in X\}$ is a collection of probability measures on $\mathfrak{B}$, where*

$$k^B(x) = k_x(B)$$

for all $B \in \mathfrak{B}$ and for all $x \in X$, then there is a unique probability measure Γ on the σ-algebra $\sigma(\mathfrak{A}, \mathfrak{B})$ of subsets of $X \times Y$ such that

$$\Gamma(A \times B) \coloneqq \int_A k^B(x)\, d\mu(x)$$

for all $A \times B \in \mathfrak{Semi}(\mathfrak{A}, \mathfrak{B})$. Furthermore, $\mu(A) = \Gamma(A \times Y)$ for all $A \in \mathfrak{A}$.

Proof. Show that Γ restricted to the product semiring $\mathfrak{Semi}(\mathfrak{A}, \mathfrak{B})$ is countably additive. Then take Γ itself to be the corresponding outer measure restricted to $\sigma(\mathfrak{A}, \mathfrak{B})$. Uniqueness follows by Proposition 2.21. Since k_x is a probability measure, it follows that $k_x(Y) = 1$. That is, the measure of the whole space is 1. And by hypothesis, $k^Y(x) = k_x(Y) = 1$ so

$$\Gamma(A \times Y) = \int_A k^Y(x)\, d\mu(x) = \int_A 1_X\, d\mu = \mu(A),$$

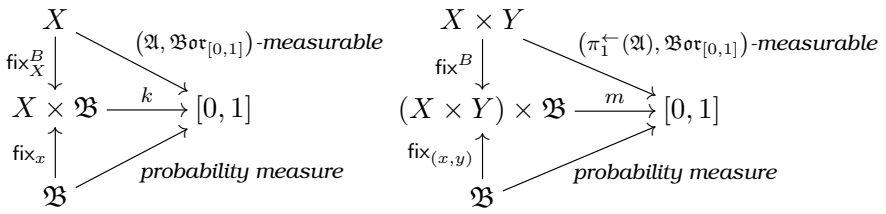
as required. ∎

Proposition 8.9. *Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces, and let $\pi_1 : X \times Y \to X$ denote the canonical projection onto the first component, with $\pi_1^{\leftarrow} : \mathfrak{A} \to \pi_1^{\leftarrow}(\mathfrak{A})$. Let $k : X \times \mathfrak{B} \to [0, 1]$ be a Markov kernel given measurable spaces $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$. If a function $m : (X \times Y) \times \mathfrak{B} \to [0, 1]$ is defined in terms of the Markov kernel k by*

$$m : (X \times Y) \times \mathfrak{B} \to [0, 1] : \big((x, y), B\big) \mapsto k(x, B),$$

then m is a Markov kernel given measurable spaces $\big(X \times Y, \pi_1^{\leftarrow}(\mathfrak{A})\big)$

and $(Y, \mathfrak{B})$*, as illustrated:*



Conversely, let $m : (X \times Y) \times \mathfrak{B} \to [0,1]$ *be a Markov kernel given measurable spaces* $\big(X \times Y, \pi_1^{\leftarrow}(\mathfrak{A})\big)$ *and* $(Y, \mathfrak{B})$*. If a function* $k : X \times \mathfrak{B} \to [0,1]$ *is defined in terms of the Markov kernel* m *by*

$$k : X \times \mathfrak{B} \to [0,1] : (x, B) \mapsto m\big((x,y), B\big),$$

for any $y \in Y$ *such that* $(x,y) \in X \times Y$*, then* k *is a Markov kernel given measurable spaces* $(X, \mathfrak{A})$ *and* $(Y, \mathfrak{B})$*.*

Proof. Let $m : (X \times Y) \times \mathfrak{B} \to [0,1]$ be defined by

$$m : (X \times Y) \times \mathfrak{B} \to [0,1] : \big((x,y), B\big) \mapsto k(x, B).$$

We want to show that m is a Markov kernel given measurable spaces $\big(X \times Y, \pi_1^{\leftarrow}(\mathfrak{A})\big)$ and $(Y, \mathfrak{B})$.

$$m \circ \mathsf{fix}^B(x,y) = m\big((x,y), B\big) = k(x, B) = k \circ \mathsf{fix}_X^B(x) = k \circ \mathsf{fix}_X^B \circ \pi_1(x,y),$$

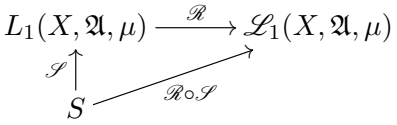
which is a composition of measurable functions. Also by definition,

$$m \circ \mathsf{fix}_{(x,y)}(B) = m\big((x,y), B\big) = k(x, B) = k \circ \mathsf{fix}_x(B),$$

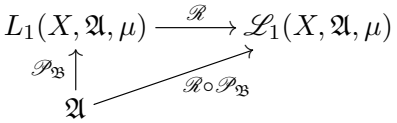
so $m \circ \mathsf{fix}_{(x,y)} = k \circ \mathsf{fix}_x$, which is a probability measure. ∎

8.3 Regular Version

Definition 8.10 (Version). Let $(X, \mathfrak{A}, \mu)$ be a measure space. Let $\mathscr{R}$ be a function $L_1(X, \mathfrak{A}, \mu) \to \mathscr{L}_1(X, \mathfrak{A}, \mu)$ which selects a representative from each equivalence class. If S is a set, and $\mathscr{S} : S \to L_1(X, \mathfrak{A}, \mu)$ is a function, then the collection $(\mathscr{R} \circ \mathscr{S})(S)$ of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions is called a **version** of the function $\mathscr{S}$.



Definition 8.11 (Regular conditional probability). Let the collection $(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}})(\mathfrak{A})$ of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions be a version of the conditional probability $\mathscr{P}_{\mathfrak{B}}$ on $\mathfrak{A}$ given $\mathfrak{B}$:



If the selection function $\mathscr{R} : L_1(X, \mathfrak{A}, \mu) \to \mathscr{L}_1(X, \mathfrak{A}, \mu)$ has the additional property that each map

$$\mathfrak{A} \to [0, 1] : A \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(A)\big)(x)$$

is a probability measure on $\mathfrak{A}$ except for possibly all x in a $\mu_{|\mathfrak{B}}{}^*$-null set, and for each $A \in \mathfrak{A}$, the map

$$X \to [0, 1] : x \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(A)\big)(x)$$

is $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, then the collection $\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(\mathfrak{A})$ of functions, or any such collection of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions without specifying $\mathscr{R}$, is called a **regular version** of the conditional probability

$\mathscr{P}_{\mathfrak{B}}$ on $\mathfrak{A}$ given $\mathfrak{B}$. If all of these $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions are modified on some fixed $\mu_{|\mathfrak{B}}{}^*$-null set, then the new collection of functions is still considered to be a regular version of $\mathscr{P}_{\mathfrak{B}}$ on $\mathfrak{A}$ given $\mathfrak{B}$. It means that any regular version of $\mathscr{P}_{\mathfrak{B}}$ on $\mathfrak{A}$ given $\mathfrak{B}$ could be redefined in such a way that the map $A \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(A)\big)(x)$ is a probability measure on all of $\mathfrak{A}$: should $A \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(A)\big)(x)$ be a probability measure on $\mathfrak{A}$ except for all x in a $\mu_{|\mathfrak{B}}{}^*$-null set N, then choose $y \in X$ with $y \notin N$, and set

$$\big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(A)\big)(x) = \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}}(A)\big)(y)$$

for all $x \in N$. Unless otherwise stated, all regular conditional probabilites will be considered so selected.

We will call a conditional probability $\mathscr{P}_{\mathfrak{B}}$ on $\mathfrak{A}$ given $\mathfrak{B}$ a **regular conditional probability** if there exists a regular version of $\mathscr{P}_{\mathfrak{B}}$.

Remark 8.12. To quote a paragraph of Loeve [5] verbatim:

> A c.pr. $P^{\mathfrak{B}}$ is said to be *regular* if, for every $A \in \mathfrak{A}$, it is possible to select $P^{\mathfrak{B}}A$ within its class of equivalence in such a manner that the $P_x^{\mathfrak{B}}$ are pr.'s on $\mathfrak{A}$ except for points x belonging to a $P_{\mathfrak{B}}$-null event N. A regular pr.f. $P^{\mathfrak{B}}$ can be said to be defined up to an equivalence, in the sense that if all the functions $P^{\mathfrak{B}}A$ are modified arbitrarily on an arbitrary but fixed $P_{\mathfrak{B}}$-null event, the new c.pr. is still regular. In particular, a regular c.pr. $P^{\mathfrak{B}}$ can be selected within its equivalence class so that $P_x^{\mathfrak{B}}$ is a pr. on $\mathfrak{A}$ for *every* $x \in X$. For example, for every x belonging to the exceptional $P_{\mathfrak{B}}$-null event N set $P_x^{\mathfrak{B}} = P_N$ where P_N is a pr. on $\mathfrak{A}$. Unless otherwise stated, regular c.pr.'s will be so selected.

Loeve continues on saying that a *regular* conditional probability is just a conditional probability which induces a particular Markov kernel; we leave it as the statement of Definition 8.11 and Example 8.14.

Discussion 8.13. Let us compare the conditional probability $P^{\mathfrak{B}}$ from Loeve in Remark 8.12 with the conditional probability $\mathscr{P}_{\mathfrak{B}}$ in Definition 8.11. The conditional probability $P^{\mathfrak{B}}$ quoted from Loeve in Remark 8.12 refers to the restriction of the operator

$$E_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$$

to the family of characteristic functions of events in $\mathfrak{A}$, which we illustrate as:

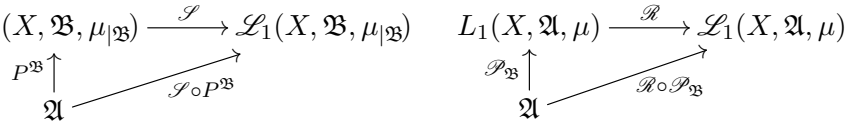
$$\begin{array}{ccc} L_1(X, \mathfrak{A}, \mu) & \xrightarrow{E_{\mathfrak{B}}} & L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}}) \\ {\scriptstyle Q\circ\chi}\uparrow & \nearrow_{P^{\mathfrak{B}}} & \\ \mathfrak{A} & & \end{array}$$

The conditional probability $\mathscr{P}_{\mathfrak{B}}$ is defined in Definition 8.1 as the composition $\mathscr{P}_{\mathfrak{B}} \circ Q \circ \chi$ in the following diagram:

$$\begin{array}{ccc} L_1(X, \mathfrak{A}, \mu) & \xrightarrow{\mathscr{E}_{\mathfrak{B}}} & L_1(X, \mathfrak{A}, \mu) \\ {\scriptstyle Q\circ\chi}\uparrow & \nearrow_{\mathscr{P}_{\mathfrak{B}}} & \\ \mathfrak{A} & & \end{array}$$

How do the regular versions of these conditional probabilities compare? They are equivalent in this sense. There exists a regular version of $P^{\mathfrak{B}}$ if and only if there exists a regular version of $\mathscr{P}_{\mathfrak{B}}$. Briefly, say $\mathscr{S}$ and $\mathscr{R}$ are selection functions (they select representatives from equivalence classes, see Definition 8.10) as in the

following diagrams:



Proposition 4.14 shows that if $\mathscr{S}\big(P^{\mathfrak{B}}(A)\big) \in \mathscr{L}_1(X,\mathfrak{B},\mu_{|\mathfrak{B}})$, then $\mathscr{S}\big(P^{\mathfrak{B}}(A)\big) \in L_1(X,\mathfrak{A},\mu)$, and we can take $\mathscr{R}\big(\mathscr{P}_{\mathfrak{B}}(A)\big) = \mathscr{S}\big(P^{\mathfrak{B}}(A)\big)$. Similarly, by Proposition 4.21, if $\mathscr{R}\big(\mathscr{P}_{\mathfrak{B}}(A)\big) \in L_1(X,\mathfrak{A},\mu)$ and is $(\mathfrak{B},\mathfrak{Bor}_{[0,1]})$-measurable, then $\mathscr{R}\big(\mathscr{P}_{\mathfrak{B}}(A)\big) \in \mathscr{L}_1(X,\mathfrak{B},\mu_{|\mathfrak{B}})$ and so we can take $\mathscr{S}\big(P^{\mathfrak{B}}(A)\big) = \mathscr{R}\big(\mathscr{P}_{\mathfrak{B}}(A)\big)$. The version in both cases is the same family of functions, it just depends on whether they are considered as residing in $\mathscr{L}_1(X,\mathfrak{B},\mu_{|\mathfrak{B}})$ or $\mathscr{L}_1(X,\mathfrak{A},\mu)$. Verify that if $\mathscr{S}\big(P^{\mathfrak{B}}(A)\big) = \mathscr{R}\big(\mathscr{P}_{\mathfrak{B}}(A)\big)$ and either of $A \mapsto \mathscr{R}\big(\mathscr{P}_{\mathfrak{B}}(A)\big)(x)$ or $A \mapsto \mathscr{S}\big(P^{\mathfrak{B}}(A)\big)(x)$ is a probability measure on $\mathfrak{A}$, then so is the other.

Example 8.14. Every regular conditional probability induces a Markov kernel; let $(X,\mathfrak{A},\mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. Let the collection $(\mathscr{R}\circ\mathscr{P}_{\mathfrak{B}})(\mathfrak{A})$ of $\mathscr{L}_1(X,\mathfrak{A},\mu)$ functions be a regular version of the conditional probability $\mathscr{P}_{\mathfrak{B}}$ on $\mathfrak{A}$ given $\mathfrak{B}$, which means by definition that the selection function $\mathscr{R} : L_1(X,\mathfrak{A},\mu) \to \mathscr{L}_1(X,\mathfrak{A},\mu)$ has the additional property that for each $A \in \mathfrak{A}$, the map

$$X \to [0,1] : x \mapsto \big[(\mathscr{R}\circ\mathscr{P}_{\mathfrak{B}})(A)\big](x)$$

is $\big(\mathfrak{B},\mathfrak{Bor}_{[0,1]}\big)$-measurable, and for each $x \in X$ the map

$$\mathfrak{A} \to [0,1] : A \mapsto \big[(\mathscr{R}\circ\mathscr{P}_{\mathfrak{B}})(A)\big](x)$$

is a probability measure on $\mathfrak{A}$. Let us define a Markov kernel

$$k : X \times \mathfrak{A} \to [0,1]$$

so that $k \circ \mathsf{fix}^A = (\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}})(A)$; that is, define

$$k : X \times \mathfrak{A} : (x, A) \mapsto \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}})(A)\big](x),$$

which we will show is a Markov kernel given measurable spaces $(X, \mathfrak{B})$ and $(X, \mathfrak{A})$.

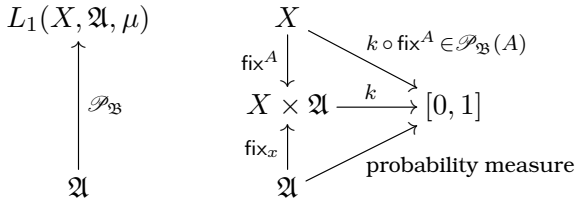
For each set $A \in \mathfrak{A}$, the map

$$k \circ \mathsf{fix}^A : X \to [0, 1] : x \mapsto k(x, A) = \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}})(A)\big](x)$$

is $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable by hypothesis, and for each element $x \in X$, the map

$$k \circ \mathsf{fix}_x : \mathfrak{A} \to [0, 1] : A \mapsto k(x, A) = \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}})(A)\big](x)$$

is a probability measure on $\mathfrak{A}$ by hypothesis, as illustrated:



This shows that k is a Markov kernel given measurable spaces $(X, \mathfrak{B})$ and $(X, \mathfrak{A})$, according to definition.

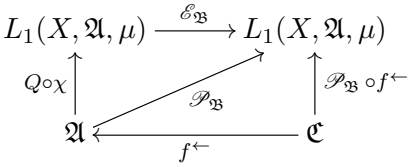
8.4 Conditional Distributions

Definition 8.15 (Conditional distribution). Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. Also let $(Y, \mathfrak{C})$ be a measurable space, and let $f : X \to Y$ be an $(\mathfrak{A}, \mathfrak{C})$-measurable

function, with $f^{\leftarrow} : \mathfrak{C} \to \mathfrak{A}$. Just as the function f induces the probability distribution $\mu \circ f^{\leftarrow}$ of f on $\mathfrak{C}$, so does the function f together with the σ-subalgebra $\mathfrak{B}$ induce the **conditional probability distribution** $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$:

$$\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow} : \mathfrak{C} \to L_1(X, \mathfrak{A}, \mu) : C \mapsto \mathscr{E}_{\mathfrak{B}} \overline{\chi_{f^{\leftarrow}(C)}}^{\mu},$$

or simply the *conditional distribution* of f on $\mathfrak{C}$ given $\mathfrak{B}$, as illustrated:



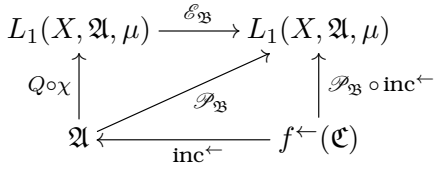
Remark 8.16. Distributions are functions induced on a σ-algebra of subsets of the codomain of a measurable function, whether the distribution is a probability distribution or a conditional distribution:

- A probability distribution is a probability measure composed with a preimage.
- A conditional distribution is a conditional probability composed with a preimage.

It is just that simple.

Discussion 8.17. You might reasonably try to avoid explicit mention of the map $f : (X, \mathfrak{A}) \to (Y, \mathfrak{B})$ in the definition of conditional distribution, perhaps by replacing the σ-algebra $\mathfrak{C}$ with $f^{\leftarrow}(\mathfrak{C})$, and

replacing $f^{\leftarrow}$ with $\text{inc}^{\leftarrow} \coloneqq \text{inclusion}^{\leftarrow}$



But one way or another, with a distribution, there is a preimage involved. There is no way around it. An $f^{\leftarrow}$ is in there somewhere.

Proposition 8.18. *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. Also let $(Y, \mathfrak{C})$ be a measurable space, and let $f : X \to Y$ be an $(\mathfrak{A}, \mathfrak{C})$-measurable function, with $f^{\leftarrow} : \mathfrak{C} \to \mathfrak{A}$. If $B \in \mathfrak{B}$ and $C \in \mathfrak{C}$, then*

$$\int_B^{\bar{}} (\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\, d\mu = \mu\big(B \cap f^{\leftarrow}(C)\big).$$

Proof.

$$\begin{aligned}
\int_B^{\bar{}} (\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\, d\mu &= \int_B^{\bar{}} \mathscr{E}_{\mathfrak{B}}\, \overline{\chi_{f^{\leftarrow}(C)}}^{\mu}\, d\mu \\
&= \int_B^{\bar{}} E_{\mathfrak{B}}\, \overline{\chi_{f^{\leftarrow}(C)}}^{\mu}\, d\mu_{|\mathfrak{B}} \\
&= \int_B \chi_{f^{\leftarrow}(C)}\, d\mu \\
&= \mu\big(B \cap f^{\leftarrow}(C)\big).
\end{aligned}$$

■

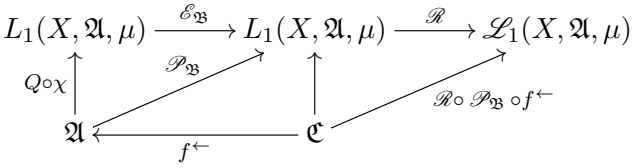
Definition 8.19 (Regular conditional distribution). Let the collection $(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(\mathfrak{C})$ of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions be a version of the conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$. If the selection function $\mathscr{R} : L_1(X, \mathfrak{A}, \mu) \to \mathscr{L}_1(X, \mathfrak{A}, \mu)$ has the additional property that the map

$$\mathfrak{C} \to [0,1] : C \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)\big)(x)$$

is a probability measure on the σ-algebra $\mathfrak{C}$ except for possibly all x in a $\mu_{|\mathfrak{B}^*}$-null set, and for each $C \in \mathfrak{C}$, the map

$$X \to [0,1] : x \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)\big)(x)$$

is $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, then the collection $\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(\mathfrak{C})$ of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions, or any such collection of $\mathscr{L}_1(X, \mathfrak{A}, \mu)$ functions without specifying $\mathscr{R}$, is called a regular version of the conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$, as illustrated:



If all of the function in the collection $(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(\mathfrak{C})$ are modified on a fixed $\mu_{|\mathfrak{B}^*}$-null set, then the new collection of functions is still considered to be a regular version of $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$. It means that any regular version of $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$ could be redefined in such a way that the map $C \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{C}} \circ f^{\leftarrow}(C)\big)(x)$ is a probability measure on all of $\mathfrak{A}$: should $C \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)\big)(x)$ be a probability measure on $\mathfrak{C}$ except for all x in a $\mu_{|\mathfrak{B}^*}$-null set N, then choose $y \in X$ with $y \notin N$, and set

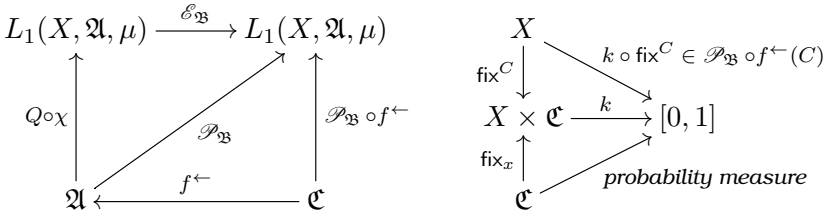
$$\big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)\big)(x) = \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)\big)(y)$$

for all $x \in N$. Unless otherwise stated, all regular conditional distributions will be considered so selected.

We will call a conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$ a **regular conditional probability distribution**, or simply a *regular conditional distribution*, if there exists a regular version of the conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$. Should the function f here happen to equal the identity function, so that $\mathfrak{C} = \mathfrak{A}$, then this definition agrees with the Definition 8.11 of a regular conditional probability, as it should be.

Remark 8.20. No conditional distribution could possibly be a probability measure since the codomain of a conditional distribution is an L_1 space rather than the interval $[0, 1]$. However, a regular conditional distribution is equivalent to a particular Markov kernel defining a whole family of probability measures, as shown in Proposition 8.21.

Proposition 8.21. *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. Also let $(Y, \mathfrak{C})$ be a measurable space, and let $f : X \to Y$ be an $(\mathfrak{A}, \mathfrak{C})$-measurable function, with $f^{\leftarrow} : \mathfrak{C} \to \mathfrak{A}$. If, given $(X, \mathfrak{B})$ and $(Y, \mathfrak{C})$, there is a Markov kernel $k : X \times \mathfrak{C} \to [0, 1]$ with $k \circ \mathsf{fix}^C \in \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)$ for each $C \in \mathfrak{C}$, then the collection $\{k \circ \mathsf{fix}^C : C \in \mathfrak{C}\}$ of $(\mathfrak{B}, \mathfrak{Bor}_{[0,1]})$-measurable functions is a regular version of the conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$, as illustrated:*



Conversely, if there exists a regular conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$, then given $(X, \mathfrak{B})$ and $(Y, \mathfrak{C})$ there is a Markov kernel $k : X \times \mathfrak{C} \to [0,1]$ with $k \circ \mathsf{fix}^C \in \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)$ for each $C \in \mathfrak{C}$.

Proof. Let $k : X \times \mathfrak{C} \to [0,1]$ be a Markov kernel given $(X, \mathfrak{B})$ and $(Y, \mathfrak{C})$, with each $k \circ \mathsf{fix}^C \in \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)$. To show that the collection $\{k \circ \mathsf{fix}^C : C \in \mathfrak{C}\}$ is a regular version of the conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$, see that we can use the hypothesis that $k \circ \mathsf{fix}^C \in \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)$ to specify the selection function:

$$\mathscr{R} : L_1(X, \mathfrak{A}, \mu) \to \mathscr{L}_1(X, \mathfrak{A}, \mu) : \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C) \mapsto k \circ \mathsf{fix}^C.$$

By the hypothesis that $k : X \times \mathfrak{C} \to [0,1]$ is a Markov kernel given $(X, \mathfrak{B})$ and $(Y, \mathfrak{C})$, this selection function has the additional properties that for each $C \in \mathfrak{C}$, the map

$$X \to [0,1] : x \mapsto \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big](x) = k \circ \mathsf{fix}^C(x)$$

is $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, and for each $x \in X$ the map

$$\mathfrak{C} \to [0,1] : C \mapsto \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big](x) = k \circ \mathsf{fix}_x(C)$$

is a probability measure on $\mathfrak{C}$, showing that $\{k \circ \mathsf{fix}^C : C \in \mathfrak{C}\}$ is a regular version of the conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ of f on $\mathfrak{C}$ given $\mathfrak{B}$ by the very definition.

Now let $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}$ be a regular conditional distribution of f on $\mathfrak{C}$ given $\mathfrak{B}$, which means we are supposing there is a selection function $\mathscr{R} : L_1(X, \mathfrak{A}, \mu) \to \mathscr{L}_1(X, \mathfrak{A}, \mu)$ with the additional properties that for each $C \in \mathfrak{C}$, the map

$$X \to [0,1] : x \mapsto \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big](x)$$

is $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, and for each $x \in X$ the map

$$\mathfrak{C} \to [0,1] : C \mapsto \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big](x)$$

is a probability measure on $\mathfrak{C}$. It means that we can define the Markov kernel $k : X \times \mathfrak{C} \to [0,1]$ given $(X, \mathfrak{B})$ and $(Y, \mathfrak{C})$ by

$$(x, C) \mapsto \big[(\mathscr{R} \circ \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big](x).$$

By definition, the selection function has the property that

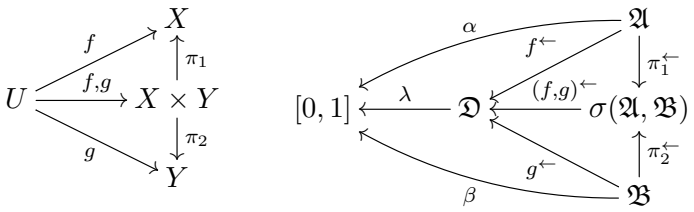
$$\mathscr{R}\big((\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big) \in (\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)$$

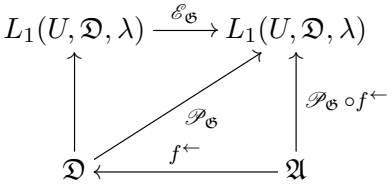
and since $k \circ \mathsf{fix}^C = \mathscr{R}\big((\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow})(C)\big)$, it follows that $k \circ \mathsf{fix}^C \in \mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow}(C)$ for each $C \in \mathfrak{C}$, as required. ■

Remark 8.22. It turns out that the proof of the following proposition using the Markov kernel approach of Proposition 8.21 is just as simple, or just as tedious, as using the definition of a regular conditional distribution. So we use the definition.

Proposition 8.23. *Let $(U, \mathfrak{D}, \lambda)$ be a probability space, with functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$ and $g : (U, \mathfrak{D}) \to (Y, \mathfrak{B})$ measurable. Set $\mathfrak{G} = g^{\leftarrow}(\mathfrak{B})$, and set the joint $\Gamma = \lambda \circ (f, g)^{\leftarrow}$, with marginal $\alpha = \Gamma \circ \pi_1^{\leftarrow}$ and marginal $\beta = \Gamma \circ \pi_2^{\leftarrow}$. If $\Gamma \ll (\alpha \times \beta)^*$ on $\sigma(\mathfrak{A}, \mathfrak{B})$, then the conditional distribution $\mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}$ of f on $\mathfrak{A}$ given $\mathfrak{G}$ is regular, and likewise, with $\mathfrak{F} = f^{\leftarrow}(\mathfrak{A})$, the conditional distribution $\mathscr{P}_{\mathfrak{F}} \circ g^{\leftarrow}$ of g on $\mathfrak{B}$ given $\mathfrak{F}$ is regular.*

Proof. Let $\Gamma \ll (\alpha \times \beta)^*$. We have set $\Gamma = \lambda \circ (f, g)^{\leftarrow}$:





In order to show that the conditional distribution $\mathscr{P}_{\mathfrak{G}}\circ f^{\leftarrow}$ of f on $\mathfrak{A}$ given $\mathfrak{G}$ is regular, by the very definition, we need to show that there is a selection function $\mathscr{R}: L_1(U,\mathfrak{D},\lambda)\to\mathscr{L}_1(U,\mathfrak{D},\lambda)$ which has the additional property that the map

$$\mathfrak{A}\to[0,1]: A\mapsto\big(\mathscr{R}\circ\mathscr{P}_{\mathfrak{G}}\circ f^{\leftarrow}(A)\big)(u)$$

is a probability measure on the σ-algebra $\mathfrak{A}$ except for possibly all u in a ${\mu_{|\mathfrak{G}}}^*$-null set, and for each $A\in\mathfrak{A}$, the map

$$U\to[0,1]: u\mapsto\big(\mathscr{R}\circ\mathscr{P}_{\mathfrak{G}}\circ f^{\leftarrow}(A)\big)(u)$$

is $\big(\mathfrak{G},\mathfrak{Bor}_{[0,1]}\big)$-measurable. And by definition, we can also modify all of the functions in the collection

$$(\mathscr{R}\circ\mathscr{P}_{\mathfrak{G}}\circ f^{\leftarrow})(\mathfrak{A})\subseteq L_1(U,\mathfrak{D},\lambda)$$

on a fixed ${\lambda_{|\mathfrak{G}}}^*$-null set, and the new collection of functions is still considered to be a regular version of $\mathscr{P}_{\mathfrak{G}}\circ f^{\leftarrow}$ of f on $\mathfrak{A}$ given $\mathfrak{G}$.

Let us provide the collection of $\big(\mathfrak{G},\mathfrak{Bor}_{[0,1]}\big)$-measurable functions,

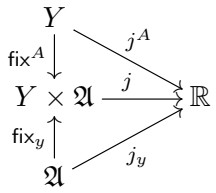
$$U\to[0,1]: u\mapsto\big(\mathscr{R}\circ\mathscr{P}_{\mathfrak{G}}\circ f^{\leftarrow}(A)\big)(u),$$

which will be a two-step process.

First, define a map $j: Y\times\mathfrak{A}\to\mathbb{R}$ as follows:

$$j: Y\times\mathfrak{A}\to\mathbb{R}: (y,A)\mapsto\int_A\frac{d\,\Gamma}{d(\alpha\times\beta)^*}(x,y)\,d\alpha(x),$$

which is illustrated in a fashion similar to a Markov kernel:



Verify that the map j makes sense; the representative function $d\Gamma/(\alpha \times \beta)^*$ is integrable with respect to the measure $\alpha \times \beta$, and this means that, by Fubini's, the iterated integrals exist and, in particular, for each $A \in \mathfrak{A}$, the map j^A defined by

$$j^A : Y \to \mathbb{R} : y \mapsto \int_A \frac{d\Gamma}{d(\alpha \times \beta)^*}(x, y)\, d\alpha(x)$$

defines an integrable function over Y with respect to the measure β.

By Proposition 3.34, each j^A can also be taken $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable. Let us show that we can redefine the codomain of each j^A to be $[0, 1]$ rather than $\mathbb{R}$.

By Proposition 5.27,

$$j^X : Y \to \mathbb{R} : y \mapsto \int_X \frac{d\Gamma}{d(\alpha \times \beta)^*}(x, y)\, d\alpha(x) = 1$$

for β^*-almost all $y \in Y$; say for all $y \in Y$ except for those y in some β^*-null subset N of Y. From the fact that each $A \subseteq X$, it follows that

$$j_y : \mathfrak{A} \to [0, 1] : A \mapsto \int_A \frac{d\Gamma}{d(\alpha \times \beta)^*}(x, y)\, d\alpha(x)$$

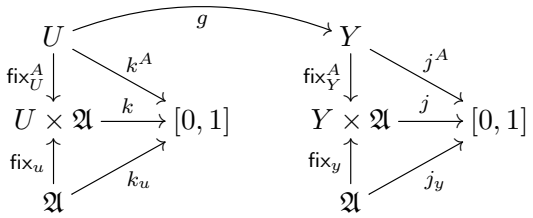
is a probability measure for all $y \in N^{\mathsf{c}}$. Let us simply redefine j^X to equal 1 on the β^*-null set N. Then we can take the codomain of all

the j^A to be $[0,1]$, with each j^A now $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, and the corresponding j_y are still probability measures for all $y \in N^{\mathsf{c}}$.

Second, we will work towards defining a map $k : U \times \mathfrak{A} \to [0,1]$. The function g is $(\mathfrak{G}, \mathfrak{B})$-measurable just by the very fact that $\mathfrak{G} = g^{\leftarrow}(\mathfrak{B})$. And since each j^A is $\big(\mathfrak{B}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, it follows that the composition

$$k^A \coloneqq j^A \circ g : U \to [0,1] : u \mapsto (j^A \circ g)(u)$$

is $\big(\mathfrak{G}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, and illustrated in the following diagram:



The function $k : U \times \mathfrak{A} \to [0,1]$ is defined in accordance with the function k^A by

$$k : U \times \mathfrak{A} \to [0,1] : (u, A) \mapsto j\big(g(u), A\big),$$

and consequently, $k_u(A) = j_{g(u)}(A)$. The required collection of $\big(\mathfrak{G}, \mathfrak{Bor}_{[0,1]}\big)$-measurable functions is the set $\{k^A : A \in \mathfrak{A}\}$.

Now let us provide the collection of probability measures

$$\mathfrak{A} \to [0,1] : A \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}(A)\big)(u).$$

We know that the j_y are probability measures on $\mathfrak{A}$, except for possibly all y in the β^*-null subset N of Y. What about the k_u? Let us argue that the k_u are also probability measures on $\mathfrak{A}$ except for

possibly all u in a $\lambda_{|\mathfrak{G}}{}^*$-null subset of U, which will turn out to be $g^{\leftarrow}(N)$.

Let us show now that $g^{\leftarrow}(N)$ is a $\lambda_{|\mathfrak{G}}{}^*$-null subset of U. We are supposing that the set N is a β^*-null subset of Y, but $\beta = \lambda \circ g^{\leftarrow}$ by Proposition 5.23, so we are supposing that N is a $\big(\lambda \circ g^{\leftarrow}\big)^*$-null set. The function g is $(\mathfrak{D}, \mathfrak{B})$-measurable by hypothesis, so by Proposition 3.11, the set $g^{\leftarrow}(N)$ is a λ^*-null subset of U. By Proposition 2.16, every set in the σ-algebra $\mathfrak{G}$ is $\lambda_{|\mathfrak{G}}{}^*$-measurable, and therefore $g^{\leftarrow}(N)$ is $\lambda_{|\mathfrak{G}}{}^*$-measurable. It follows then by Proposition 3.14 that

$$\lambda_{|\mathfrak{G}}{}^*\big(g^{\leftarrow}(N)\big) = \lambda^*\big(g^{\leftarrow}(N)\big) = 0,$$

saying that $g^{\leftarrow}(N)$ is a $\lambda_{|\mathfrak{G}}{}^*$-null subset of U.

Finally, the k_u are probability measures for all u off the $\lambda_{|\mathfrak{G}}{}^*$-null subset $g^{\leftarrow}(N)$ of U since they are equal to the $j_{g(u)}$, which we know are probability measures, at least for all $g(u)$ off N, or equivalently, for all u off $g^{\leftarrow}(N)$. The required collection of probability measures is $\{k_u : u \notin g^{\leftarrow}(N)\}$.

We have shown that there is a selection function

$$\mathscr{R} : L_1(U, \mathfrak{D}, \lambda) \to \mathscr{L}_1(U, \mathfrak{D}, \lambda)$$

defined on the range of $\mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}$ by

$$\mathscr{R}\big(\mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}(A)\big) = k^A$$

which has the additional property that each map

$$\mathfrak{A} \to [0,1] : A \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}(A)\big)(u) = k_u(A)$$

is a probability measure on the σ-algebra $\mathfrak{A}$ except for possibly all u in the $\lambda_{|\mathfrak{G}}{}^*$-null set $g^{\leftarrow}(N)$, and for each $A \in \mathfrak{A}$, the map

$$U \to [0,1] : u \mapsto \big(\mathscr{R} \circ \mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}(A)\big)(u) = k^A(u)$$

is $(\mathfrak{G}, \mathfrak{Bor}_{[0,1]})$-measurable. Therefore, the conditional distribution $\mathscr{P}_{\mathfrak{G}} \circ f^{\leftarrow}$ of f on $\mathfrak{A}$ given $\mathfrak{G}$ is regular, as required. ■

8.5 Conditional Independence

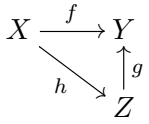
The sole purpose for this section is to support Proposition 8.31. Unfortunately, the notation I am trying to use here has gotten so bad that I just cannot stand it. It causes mental instability. Just look at it:

- $\lambda_{f|h}$
- $\lambda_{f|h^{\leftarrow}(z)}$
- $\lambda_{f|h=z}$

Whenever you get that much garbage in a subscript, you need to seriously take a long break and rethink what you are trying to do. There is plenty of therapeutic reading on how to deal with truly gross subscripts.

This whole section needs a going over.

Proposition 8.24. *Let a function $f : X \to Y$ factor as $g \circ h$, with $h : X \to Z$ and $g : Z \to Y$, as illustrated:*



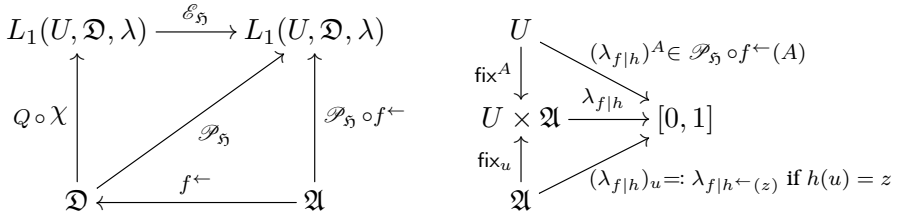
For any $z \in Z$, the value of f on the preimage $h^{\leftarrow}(z)$ is constant, and is equal to $g(z)$.

Proof. Let $x \in h^{\leftarrow}(z)$. Then $h(x) = z$, so $f(x) = (g \circ h)(x) = g(h(x)) = g(z)$. ■

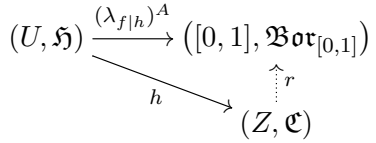
Discussion 8.25. The idea here is to identify a particular probablity measure, to be denoted $\lambda_{f|h^{\leftarrow}(z)}$, given a Markov kernel. Let $(U, \mathfrak{D}, \lambda)$ be a probability space, with functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$ and $h : (U, \mathfrak{D}) \to (Z, \mathfrak{C})$ both measurable. We will set $\mathfrak{H} = h^{\leftarrow}(\mathfrak{C})$ because subscripts do not suffer symbols like $h^{\leftarrow}(\mathfrak{C})$ or $\sigma(h)$ well. Also let the conditional distribution

$$\mathscr{P}_{\mathfrak{H}} \circ f^{\leftarrow} : \mathfrak{A} \to L_1(U, \mathfrak{D}, \lambda)$$

of f on $\mathfrak{A}$ given $\mathfrak{H}$ be regular, and let $\lambda_{f|h}$ denote the corresponding Markov kernel, as illustrated:



For each set $A \in \mathfrak{A}$, the function $(\lambda_{f|h})^A$ factors as some regression function r composed with h, that is, $(\lambda_{f|h})^A = r \circ h$, which was the heart of Proposition 4.34:



So for any $z \in Z$, the value of the function $(\lambda_{f|h})^A$ on the preimage $h^{\leftarrow}(z)$ is constant and is equal to $r(z)$, as was just shown in Proposition 8.24. And by the very definition of the Markov kernel itself:

$$(\lambda_{f|h})_u(A) = \lambda_{f|h}(u, A) = (\lambda_{f|h})^A(u).$$

This implies that the value of the measure $(\lambda_{f|h})_u$ at A is the same for all $u \in U$ in the preimage $h^{\leftarrow}(z)$. We identify the measure $\lambda_{f|h^{\leftarrow}(z)}$ on $\mathfrak{A}$ by

$$\lambda_{f|h^{\leftarrow}(z)} \coloneqq (\lambda_{f|h})_u \quad \text{if } u \in h^{\leftarrow}(z).$$

The value of this measure on $A \in \mathfrak{A}$ relates to the regression function:

$$\lambda_{f|h^{\leftarrow}(z)}(A) = r(z) \quad \text{if } h(u) = z.$$

There has just got to be better notation than $\lambda_{f|h^{\leftarrow}(z)}$. Worse would be $\lambda_{f|h=z}$, so we are not going in that direction.

Proposition 8.26 (Joint decomposition). *Let $(U, \mathfrak{D}, \lambda)$ be a probability space, with functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$ and $h : (U, \mathfrak{D}) \to (Z, \mathfrak{C})$ measurable, and set $\mathfrak{H} = h^{\leftarrow}(\mathfrak{C})$. If the conditional distribution*

$$\mathscr{P}_{\mathfrak{H}} \circ f^{\leftarrow} : \mathfrak{A} \to L_1(U, \mathfrak{D}, \lambda)$$

of f on $\mathfrak{A}$ given $\mathfrak{H}$ is regular, and if

$$k \in L_1\big(X \times Z, \sigma(\mathfrak{A}, \mathfrak{C}), \lambda_{f,h}\big),$$

then

$$\int_{X\times Z} k \, d\lambda_{f,h} = \int_Z \int_X k(x,z) \, d\lambda_{f|h^{\leftarrow}(z)}(x) \, d\lambda_h(z).$$

Proof. Let us show that equality holds for the function $k = \chi_{A\times C}$,

where $A \in \mathfrak{A}$ and $C \in \mathfrak{C}$:

$$
\begin{aligned}
&\int_{X\times Z} \chi_{A\times C}\, d\lambda_{f,h} \\
&= \lambda(f^{\leftarrow}(A) \cap h^{\leftarrow}(C)) \\
&= \int_U \chi_{f^{\leftarrow}(A)} \cdot \chi_{h^{\leftarrow}(C)}\, d\lambda \\
&= \int_U \mathscr{E}_{\mathfrak{H}}\big(\chi_{f^{\leftarrow}(A)} \cdot \chi_{h^{\leftarrow}(C)}\big)\, d\lambda \\
&= \int_U \chi_{h^{\leftarrow}(C)} \cdot \mathscr{E}_{\mathfrak{H}}\big(\chi_{f^{\leftarrow}(A)}\big)\, d\lambda && (\mathfrak{H}, \mathfrak{Bor}_{\mathbb{R}})\text{-measurable} \\
&= \int_U \chi_{h^{\leftarrow}(C)} \cdot (r \circ h)\, d\lambda && \mathscr{E}_{\mathfrak{H}}\big(\chi_{f^{\leftarrow}(A)}\big) = r \circ h \\
&= \int_U \big((\chi_C \cdot r) \circ h\big)\, d\lambda && \chi_{h^{\leftarrow}(C)} = \chi_C \circ h \\
&= \int_Z (\chi_C \cdot r)\, d\lambda_h && \text{change of variable} \\
&= \int_Z \chi_C(z) \cdot \lambda_{f|h^{\leftarrow}(z)}(A)\, d\lambda_h(z) && \lambda_{f|h^{\leftarrow}(z)}(A) = r(z) \text{ if } h(u) = z \\
&= \int_Z \chi_C(z) \left(\int_X \chi_A(x)\, d\lambda_{f|h^{\leftarrow}(z)}(x)\right) d\lambda_h(z) \\
&= \int_Z \int_X \chi_A(x) \cdot \chi_C(z)\, d\lambda_{f|h^{\leftarrow}(z)}(x)\, d\lambda_h(z),
\end{aligned}
$$

as required. ■

Proposition 8.27 (Conditional density). *Let $(U, \mathfrak{D}, \lambda)$ be a probability space, with functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$ and $h : (U, \mathfrak{D}) \to (Z, \mathfrak{C})$ measurable. Let μ and ρ be σ-finite measures on respective σ-*

algebras $\mathfrak{A}$ and $\mathfrak{C}$. Also let $\lambda_{f,h} \ll \mu \times \rho$ on $\sigma(\mathfrak{A}, \mathfrak{C})$, and set

$$m_{f,h} = \frac{d\lambda_{f,h}}{d(\mu \times \rho)^*}$$

and

$$m_h = \frac{d\lambda_h}{d\rho}.$$

Then the measure $\lambda_{f|h^{\leftarrow}(z)}$ on $\mathfrak{A}$ has the following density function with respect to the measure μ:

$$\frac{d\lambda_{f|h^{\leftarrow}(z)}}{d\mu} : X \to \mathbb{R} : x \mapsto \frac{m_{f,h}(x,z)}{m_h(z)}.$$

Proof. It is sufficient to show that

$$\lambda_{f|h^{\leftarrow}(z)}(A) = \int_A \frac{m_{f,h}(x,z)}{m_h(z)}\, d\mu(x) \quad \text{for all } A \in \mathfrak{A}.$$

We will start by showing that if

$$k \in L_1\big(X \times Z, \sigma(\mathfrak{A}, \mathfrak{C}), \lambda_{f,h}\big),$$

then

$$\int_{X\times Z} k\, d\lambda_{f,h} = \int_Z \int_X k(x,z) \frac{m_{f,h}(x,z)}{m_h(z)}\, d\mu(x)\, d\lambda_h(z),$$

for then by Proposition 8.26, it will follow that

$$\int_X k(x,z) \frac{m_{f,h}(x,z)}{m_h(z)}\, d\mu(x) = \int_X k(x,z)\, d\lambda_{f|h^{\leftarrow}(z)}(x),$$

and by setting $k(x,z) = \chi_A(x)$, we will obtain the result.

Let $\lambda_{f,h} \ll \mu \times \rho$ on $\sigma(\mathfrak{A}, \mathfrak{C})$, and let $k \in L_1\big(X \times Z, \sigma(\mathfrak{A}, \mathfrak{C}), \lambda_{f,h}\big)$. Then

$$\begin{aligned}
&\int_{X\times Z} k \, d\lambda_{f,h} \\
&= \int_{X\times Z} k(x,z) m_{f,h}(x,z) \, d(\mu\times\rho)(x,z) && m_{f,h} = \frac{d\lambda_{f,h}}{d(\mu\times\rho)} \\
&= \int_Z \int_X k(x,z) \frac{m_{f,h}(x,z)}{m_h(z)} m_h(z) \, d\mu(x) \, d\rho(z) && \text{well-chosen one} \\
&= \int_Z \int_X k(x,z) \frac{m_{f,h}(x,z)}{m_h(z)} \, d\mu(x) \, d\lambda_h(z) && m_h = \frac{d\lambda_h}{d\rho}
\end{aligned}$$

By Proposition 8.26,

$$\int_{X\times Z} k \, d\lambda_{f,h} = \int_Z \int_X k(x,z) \, d\lambda_{f|h^{\leftarrow(z)}}(x) \, d\lambda_h(z).$$

By transitivity of equality,

$$\int_Z \int_X k(x,z) \frac{m_{f,h}(x,z)}{m_h(z)} \, d\mu(x) \, d\lambda_h(z) = \int_Z \int_X k(x,z) \, d\lambda_{f|h^{\leftarrow(z)}}(x) \, d\lambda_h(z).$$

It follows that

$$\int_X k(x,z) \frac{m_{f,h}(x,z)}{m_h(z)} \, d\mu(x) = \int_X k(x,z) \, d\lambda_{f|h^{\leftarrow(z)}}(x).$$

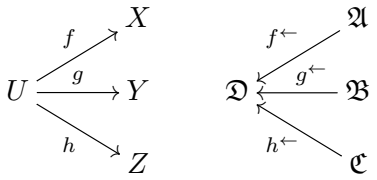
Set $k(x,z) = \chi_A(x)$ and apply:

$$\int_A \frac{m_{f,h}(x,z)}{m_h(z)} \, d\mu(x) = \int_A d\lambda_{f|h^{\leftarrow(z)}}(x) = \lambda_{f|h^{\leftarrow(z)}}(A),$$

as required. ■

Definition 8.28. Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. Also, let $\mathfrak{C}$ and $\mathfrak{D}$ be σ-subalgebras of $\mathfrak{A}$. Then $\mathfrak{C}$ and $\mathfrak{D}$ are **conditionally independent σ-subalgebras** given $\mathfrak{B}$ if $\mathscr{P}_{\mathfrak{B}}(C \cap D) = \mathscr{P}_{\mathfrak{B}}(C)\,\mathscr{P}_{\mathfrak{B}}(D)$ for all $C \in \mathfrak{C}$ and for all $D \in \mathfrak{D}$. The notation $\mathfrak{C} \perp\!\!\!\perp \mathfrak{D} \mid \mathfrak{B}$ may be used.

Notation 8.29. Let $(U, \mathfrak{D}, \lambda)$ be a probability space, and let functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$, $g : (U, \mathfrak{D}) \to (Y, \mathfrak{B})$, and $h : (U, \mathfrak{D}) \to (Z, \mathfrak{C})$ be measurable:



Say that f and g are conditionally independent given h if

$$f^{\leftarrow}(\mathfrak{A}) \perp\!\!\!\perp g^{\leftarrow}(\mathfrak{B}) \mid h^{\leftarrow}(\mathfrak{C}).$$

The notation $f \perp\!\!\!\perp g \mid h$ may be used.

Notation 8.30. Let $(U, \mathfrak{D}, \lambda)$ be a probability space, with functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$, $g : (U, \mathfrak{D}) \to (Y, \mathfrak{B})$ and $h : (U, \mathfrak{D}) \to (Z, \mathfrak{C})$ measurable. Let μ, ν, and ρ be σ-finite measures on $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$,

respectively, and let $\lambda_{f,g,h} \ll (\mu \times \nu \times \rho)^*$ on $\sigma(\mathfrak{A}, \mathfrak{B}, \mathfrak{C})$. Set:

$$\begin{aligned} j &= \frac{d\lambda_{f,g,h}}{d(\mu \times \nu \times \rho)^*} : X \times Y \times Z \to \mathbb{R}, \\ m_{f,h} &= \frac{d\lambda_{f,h}}{d(\mu \times \rho)^*} : X \times Z \to \mathbb{R}, \\ m_{g,h} &= \frac{d\lambda_{g,h}}{d(\nu \times \rho)^*} : Y \times Z \to \mathbb{R}, \\ m_h &= \frac{d\lambda_h}{d\rho} : Z \to \mathbb{R}. \end{aligned}$$

The letter 'j' was intended to roughly correspond with 'joint density,' and likewise the letter 'm' with 'marginal density.' It was shown back in Proposition 5.24 that if $\lambda_{f,g,h} \ll \mu \times \nu \times \rho$, then all four Radon-Nikodym derivatives displayed here exist. Actually, that was shown for the case where only two functions f and g were involved. Same proof. Recall

$$\begin{aligned} m_{f,h}(x,z) &= \int_Y j(x,y,z)\, d\nu(y), \\ m_{g,h}(y,z) &= \int_X j(x,y,z)\, d\mu(x), \\ m_h(z) &= \int_{X \times Y} j(x,y,z)\, d(\mu \times \nu)(x,y) \end{aligned}$$

Also, for the following ratios of Radon-Nikodym derivatives, let:

$$\begin{aligned} p(x, y \mid z) &= \frac{j(x, y, z)}{m_h(z)}, \\ p(x \mid z) &= \frac{m_{f,h}(x, z)}{m_h(z)}, \\ p(y \mid z) &= \frac{m_{g,h}(y, z)}{m_h(z)}, \\ p(x \mid y, z) &= \frac{j(x, y, z)}{m_{g,h}(y, z)}, \end{aligned}$$

and so forth.

Proposition 8.31. *Let $(U, \mathfrak{D}, \lambda)$ be a probability space, and let functions $f : (U, \mathfrak{D}) \to (X, \mathfrak{A})$, $g : (U, \mathfrak{D}) \to (Y, \mathfrak{B})$, and $h : (U, \mathfrak{D}) \to (Z, \mathfrak{C})$ be measurable, with σ-finite measures μ, ν, and ρ on respective σ-algebras $\mathfrak{A}$, $\mathfrak{B}$, and $\mathfrak{C}$. Also let $\lambda_{f,g,h} \ll \mu \times \nu \times \rho$. The following are equivalent:*

1. $f \perp\!\!\!\perp g \mid h$,
2. $\lambda_{f,g|h^{\leftarrow}(z)}(A \times B) = \lambda_{f|h^{\leftarrow}(z)}(A) \cdot \lambda_{g|h^{\leftarrow}(z)}(B)$ *for all* $A \in \mathfrak{A}$, $B \in \mathfrak{B}$, *and ρ-almost all* $z \in Z$,
3. $p(x, y \mid z) = p(x \mid z) \cdot p(y \mid z)$,
4. $p(x \mid y, z) = p(x \mid z)$.

Proof. Set $\mathfrak{F} = f^{\leftarrow}(\mathfrak{A})$, $\mathfrak{G} = g^{\leftarrow}(\mathfrak{B})$, and $\mathfrak{H} = h^{\leftarrow}(\mathfrak{C})$.
(1) $\Leftrightarrow$ (2)
Let $f \perp\!\!\!\perp g \mid h$. By the very definition of conditionally independent σ-subalgebras, this is saying nothing but

$$\mathscr{P}_{\mathfrak{H}}\big(f^{\leftarrow}(A) \cap g^{\leftarrow}(B)\big) = \mathscr{P}_{\mathfrak{H}}\big(f^{\leftarrow}(A)\big) \cdot \mathscr{P}_{\mathfrak{H}}\big(g^{\leftarrow}(B)\big) \tag{8.1}$$

for all $A \in \mathfrak{A}$ and for all $B \in \mathfrak{B}$. This equality can be written as

$$\big(\mathscr{P}_{\mathfrak{H}} \circ (f,g)^{\leftarrow}\big)(A \times B) = (\mathscr{P}_{\mathfrak{H}} \circ f^{\leftarrow})(A) \cdot (\mathscr{P}_{\mathfrak{H}} \circ g^{\leftarrow})(B). \tag{8.2}$$

(2) $\Leftrightarrow$ (3)
By Proposition 8.27,

$$\begin{aligned}
\lambda_{f,g|h^{\leftarrow}(z)}(A \times B) &= \int_{A\times B} p(x,y \mid z)\, d(\mu \times \nu)(x,y),\\
\lambda_{f|h^{\leftarrow}}(z)(A) &= \int_{A} p(x \mid z)\, d\mu(x),\\
\lambda_{g|h^{\leftarrow}}(z)(B) &= \int_{B} p(y \mid z)\, d\nu(y).
\end{aligned}$$

By (2) then,

$$\begin{aligned}
\int_{A\times B} p(x,y \mid z)\, d(\mu \times \nu)(x,y) &= \int_{A} p(x \mid z)\, d\mu(x) \int_{B} p(y \mid z)\, d\nu(y)\\
&= \int_{A\times B} p(x \mid z) \cdot p(y \mid z)\, d(\mu \times \nu)(x,y).
\end{aligned}$$

(3) $\Leftrightarrow$ (4)
Divide both sides of (3) by $p(y \mid z)$.

■

9 Bayesian Statistics

STATISTICS is a mathematical analysis of guesswork. This kind of guesswork is called inductive inference. Its purpose is to identify a probability distribution or a conditional distribution. The particular kind of guesswork we will use is considered *Bayesian* since it involves a parameter that induces a probability distribution. A parameter here is just a measurable function.

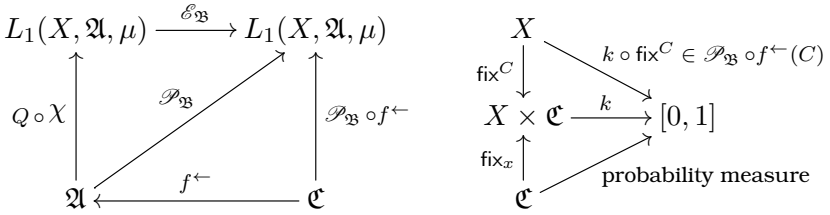
There are no topological restrictions imposed on the objects in this chapter. No Borel space conditions, no separable conditions. We rely instead upon measure-theoretic conditions.

The following preface provides the smallest possible glimpse of Bayesian statistics. It is more an article of faith. The details can be found in 9.2 Details.

9.1 Preface

Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{C})$ be measurable spaces. Suppose $(X, \mathfrak{A}, \mu)$ is a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The conditional probability $\mathscr{P}_{\mathfrak{B}} : \mathfrak{A} \to L_1(X, \mathfrak{A}, \mu)$ given $\mathfrak{B}$ could not possibly be a probability measure since the codomain of $\mathscr{P}_{\mathfrak{B}}$ is quotient space $L_1(X, \mathfrak{A}, \mu)$ rather than the interval $[0, 1]$. For the same reason, no conditional distribution $\mathscr{P}_{\mathfrak{B}} \circ f^{\leftarrow} : \mathfrak{C} \to L_1(X, \mathfrak{A}, \mu)$ given $\mathfrak{B}$ could possibly be a probability measure, with $f : X \to Y$ here being any $(\mathfrak{A}, \mathfrak{C})$-measurable function. However, a *regular* conditional

distribution is equivalent to a particular Markov kernel defining a whole family of probability measures on the σ-algebra $\mathfrak{C}$:

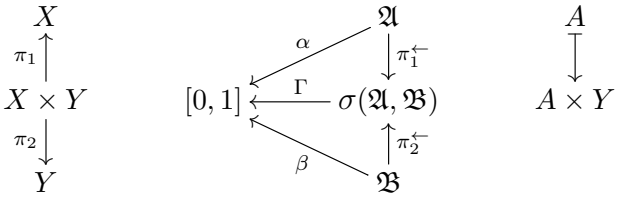


Part of the assumption that k is a Markov kernel includes the assumption that each $k\circ\mathsf{fix}^C$ is $\big(\mathfrak{A},\mathfrak{Bor}_{[0,1]}\big)$-measurable. By some schools of thought, a family of probability measures itself, indexed by a parameter space, is considered to be a statistical experiment.

Let

$$\big(X\times Y,\sigma(\mathfrak{A},\mathfrak{B}),\Gamma\big)$$

be a probability space. The measure Γ is called the *joint*. Call the measurable space $(X,\mathfrak{A})$ the *attribute space*, or *parameter space*, and call the measurable space $(Y,\mathfrak{B})$ the *sample space*. Let π_1 and π_2 denote the respective coordinate projections of $X\times Y$ onto X and Y. Also let $\pi_1^{\leftarrow}$ and $\pi_2^{\leftarrow}$ denote the respective induced maps $\pi_1^{\leftarrow}:\mathfrak{A}\to\sigma(\mathfrak{A},\mathfrak{B})$ and $\pi_2^{\leftarrow}:\mathfrak{B}\to\sigma(\mathfrak{A},\mathfrak{B})$, as illustrated:



The projection π_1 is $\big(\mathfrak{A},\sigma(\mathfrak{A},\mathfrak{B})\big)$-measurable, and induces the marginal probability distribution $\alpha\coloneqq\Gamma\circ\pi_1^{\leftarrow}$ of π_1 on the parameter

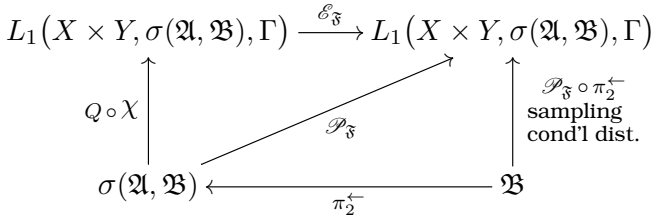
space σ-algebra $\mathfrak{A}$, which is called the *prior probability distribution*. The projection π_2 is $\big(\mathfrak{B}, \sigma(\mathfrak{A}, \mathfrak{B})\big)$-measurable, and induces the marginal probability distribution $\beta \coloneqq \Gamma \circ \pi_2^{\leftarrow}$ of π_2 on the sample space σ-algebra $\mathfrak{B}$, which is called the *predictive probability distribution*. Consequently:

$$\alpha(A) = (\Gamma \circ \pi_1^{\leftarrow})(A) = \Gamma(A \times Y) \quad \text{for all } A \in \mathfrak{A},$$

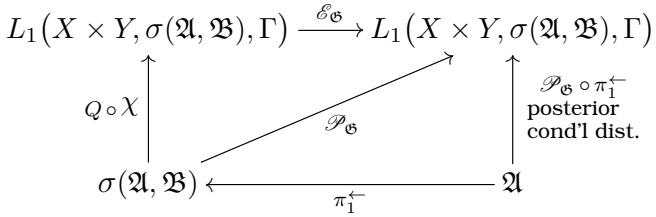
and

$$\beta(B) = (\Gamma \circ \pi_2^{\leftarrow})(B) = \Gamma(X \times B) \quad \text{for all } B \in \mathfrak{B}.$$

Set $\mathfrak{F} = \pi_1^{\leftarrow}(\mathfrak{A})$. The conditional distribution $\mathscr{P}_{\mathfrak{F}} \circ \pi_2^{\leftarrow}$ of π_2 on $\mathfrak{B}$ given $\mathfrak{F}$ is called the **sampling conditional distribution**, as illustrated:



Set $\mathfrak{G} = \pi_2^{\leftarrow}(\mathfrak{B})$. The conditional distribution $\mathscr{P}_{\mathfrak{G}} \circ \pi_1^{\leftarrow}$ of π_1 on $\mathfrak{A}$ given $\mathfrak{G}$ is called the **posterior conditional distribution**, as illustrated:



The sampling and posterior conditional distributions could not possibly be probability measures, and this should be clear since their codomains are L_1 spaces rather than the interval $[0,1]$. However, were they *regular* conditional distributions, then they would induce whole families of probability measures on the parameter and sample space σ-algebras. What measure-theoretic conditions might be sufficient to imply that these conditional distributions are regular?

One simple condition is when $\Gamma \ll (\alpha \times \beta)^*$. Then both the sampling conditional distribution and the posterior conditional distribution are regular. This is just an application of Proposition 8.23.

Here, in turn, is a measure-theoretic condition sufficient to imply that $\Gamma \ll (\alpha \times \beta)^*$. Suppose $\{P_x : x \in X\}$ is a family of probability measures on the sample space σ-algebra $\mathfrak{B}$, and suppose that the σ-algebra $\mathfrak{A}$ is such that for each $B \in \mathfrak{B}$, the function

$$P^B : X \to [0,1] : x \mapsto P_x(B)$$

is $\big(\mathfrak{A}, \mathfrak{Bor}_{[0,1]}\big)$-measurable. Further suppose the joint probability measure Γ is defined in terms of a given probability measure γ on the parameter space σ-algebra $\mathfrak{A}$ by

$$\Gamma(A \times B) = \int_A P^B(x)\, d\gamma(x) \quad \text{for all } B \in \mathfrak{B} \text{ and all } A \in \mathfrak{A}.$$

Then it turns out that γ must equal α. If also ν is a σ-finite measure on $\mathfrak{B}$ such that each $P_x \ll \nu$, and if $f : X \times Y \to \mathbb{R}$ is a $\big(\sigma(\mathfrak{A}, \mathfrak{B}), \mathfrak{Bor}_{\mathbb{R}}\big)$-measurable function such that each

$$f_x = dP_x/d\nu, \tag{9.1}$$

then $\Gamma \ll (\alpha \times \nu)^*$ with $f = d\Gamma/d(\alpha \times \nu)^*$, and consequently $\Gamma \ll (\alpha \times \beta)^*$. Further, if the function $g : X \times Y \to \mathbb{R}$ is defined by setting

$$g(x, y) = \frac{f(x, y)}{\int_X f(x, y)\, d\alpha(x)},$$

then $g = d\Gamma/d(\alpha \times \beta)^*$. This implies that $\beta \ll \nu$ on $\mathfrak{B}$, for if $B \in \mathfrak{B}$, then

$$\beta(B) = \Gamma(X \times B) = \int_{X \times B} f\, d(\alpha \times \nu) = \int_B \left(\int_X f(x, y)\, d\alpha(x) \right) d\nu(y).$$

This says

$$\frac{d\beta}{d\nu} : Y \to \mathbb{R} : y \mapsto \int_X f(x, y)\, d\alpha(x).$$

The family $\{P_x : x \in X\}$ of probability measures on the sample space σ-algebra $\mathfrak{B}$ must satisfy

$$P_x(B) = \int_B g(x, y)\, d\beta(y) \quad \text{for all } B \in \mathfrak{B},$$

and a family $\{\alpha_y : y \in Y\}$ of probability measures on the parameter space σ-algebra $\mathfrak{A}$, defined for those $y \in Y$ such that f^y is integrable with respect to α, is defined by

$$\alpha_y(A) = \int_A g(x, y)\, d\alpha(x) \quad \text{for all } A \in \mathfrak{A}, \tag{9.2}$$

or equivalently by

$$\frac{d\alpha_y}{d\alpha} : X \to \mathbb{R} : x \mapsto \frac{f(x, y)}{\int_X f(x, y)\, d\alpha(x)}.$$

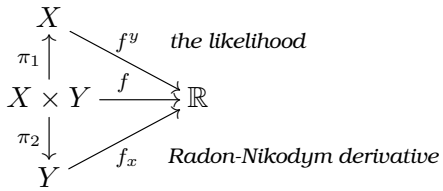
The equality (9.2) may also be seen as Bayes' Theorem in the case where there is a σ-finite measure λ on $\mathfrak{A}$ with $\alpha \ll \lambda$, say with density function $h = d\alpha/d\lambda \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$. In this case,

$$\alpha_y(A) = \frac{\displaystyle\int_A f(x,y)h(x)\,d\lambda(x)}{\displaystyle\int_X f(x,y)h(x)\,d\lambda(x)} \quad \text{for all } A \in \mathfrak{A},$$

or equivalently,

$$\begin{aligned}\frac{d\alpha_y}{d\lambda} : X \to \mathbb{R} : x \mapsto & \frac{f(x,y)h(x)}{\displaystyle\int_X f(x,y)h(x)\,d\lambda(x)} \\ &= \frac{f^y(x)h(x)}{\displaystyle\int_X f^y(x)h(x)\,d\lambda(x)},\end{aligned}$$

where $f(x,y)$, or in particular f^y, might be referred to as "the likelihood," illustrated in the following diagram:

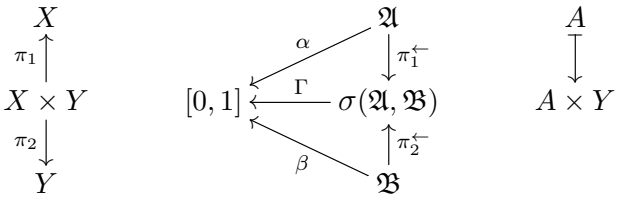


The collection $\{f^y : y \in Y\}$ of likelihood functions $X \to \mathbb{R}$ is indexed by the sample space Y, which is one way that a particular sample y may be used in applications; the sample itself determines the likelihood function by indexing into the collection. It is interesting how the sample, or **observation**, or **data**, can be seen

as selecting a likelihood function from a collection, and warrants further thought. On the other hand, you always need information to make an intelligent selection. Perhaps that is all there is to it.

9.2 Details

Terminology 9.1. Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces. Let $\big(X \times Y, \sigma(\mathfrak{A}, \mathfrak{B}), \Gamma\big)$ be a probability space. The measure Γ is called the **joint probability measure**, or simply the *joint*. Call the measurable space $(X, \mathfrak{A})$ the **attribute space**, or **parameter space**, and call the measurable space $(Y, \mathfrak{B})$ the **sample space**. Let π_1 and π_2 denote the respective coordinate projections of $X \times Y$ onto X and Y. Also let $\pi_1^{\leftarrow}$ and $\pi_2^{\leftarrow}$ denote the respective induced maps $\pi_1^{\leftarrow} : \mathfrak{A} \to \sigma(\mathfrak{A}, \mathfrak{B})$ and $\pi_2^{\leftarrow} : \mathfrak{B} \to \sigma(\mathfrak{A}, \mathfrak{B})$, as illustrated:



The projection π_1 is $\big(\mathfrak{A}, \sigma(\mathfrak{A}, \mathfrak{B})\big)$-measurable and induces the marginal probability distribution $\Gamma \circ \pi_1^{\leftarrow}$ of π_1 on the parameter space σ-algebra $\mathfrak{A}$, which is called the **prior probability distribution**, or simply the *prior*, and which we will denote by α. The projection π_2 is $\big(\mathfrak{B}, \sigma(\mathfrak{A}, \mathfrak{B})\big)$-measurable and induces the marginal probability distribution $\Gamma \circ \pi_2^{\leftarrow}$ of π_2 on the sample space σ-algebra $\mathfrak{B}$, which is called the **predictive probability distribution**, or simply the *predictive*, and which we will denote by β. Consequently:

$$\alpha(A) = (\Gamma \circ \pi_1^{\leftarrow})(A) = \Gamma(A \times Y) \quad \text{for all } A \in \mathfrak{A},$$

and

$$\beta(B) = (\Gamma \circ \pi_2^{\leftarrow})(B) = \Gamma(X \times B) \quad \text{for all } B \in \mathfrak{B}.$$

Notation 9.2. Whenever X, Y, and Z are sets and $f : X \times Y \to Z$ is a function, we will let f^y denote the map

$$f^y : X \to Z : x \mapsto f^y(x) = f(x, y),$$

and we will let f_x denote the map

$$f_x : Y \to Z : y \mapsto f_x(y) = f(x, y).$$

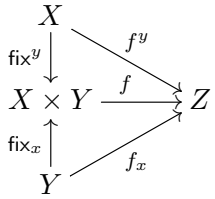
Also, for each $y \in Y$ let fix^y denote the map

$$\mathsf{fix}^y : X \to X \times Y : x \mapsto (x, y)$$

and for each $x \in X$ let fix_x denote the map

$$\mathsf{fix}_x : Y \to X \times Y : y \mapsto (x, y)$$

as illustrated:



Proposition 9.3. *Let $(X, \mathfrak{A})$ and $(Y, \mathfrak{B})$ be measurable spaces. Let $\{P_x : x \in X\}$ be a family of probability measures on the σ-algebra $\mathfrak{B}$, and let the σ-algebra $\mathfrak{A}$ be such that for each $B \in \mathfrak{B}$, the function*

$$P^B : X \to [0, 1] : x \mapsto P_x(B)$$

is $(\mathfrak{A}, \mathfrak{Bor}_{[0,1]})$*-measurable. Also let the joint probability measure* Γ *be defined in terms of a given probability measure* γ *on the* σ*-algebra* $\mathfrak{A}$ *by*

$$\Gamma(A \times B) = \int_A P^B(x)\, d\gamma(x) \quad \textit{for all } B \in \mathfrak{B} \textit{ and all } A \in \mathfrak{A}.$$

If ν *is a* σ*-finite measure on* $\mathfrak{B}$ *such that each* $P_x \ll \nu$*, and if* $f : X \times Y \to \mathbb{R}$ *is a* $(\sigma(\mathfrak{A}, \mathfrak{B}), \mathfrak{Bor}_{\mathbb{R}})$*-measurable function such that each*

$$f_x = dP_x/d\nu,$$

then $\Gamma \ll (\alpha \times \nu)^*$ *and* $f = d\Gamma/d(\alpha \times \nu)^*$*, and consequently* $\Gamma \ll (\alpha \times \beta)^*$*. Define a function* $g : X \times Y \to \mathbb{R}$ *by*

$$g : X \times Y \to \mathbb{R} : (x, y) \mapsto \frac{f(x, y)}{\int_X f(x, y)\, d\alpha(x)}.$$

Then $g = d\Gamma/d(\alpha \times \beta)^*$*.*

Proof. Let ν be a σ-finite measure on $\mathfrak{B}$ such that each $P_x \ll \nu$, and let $f : X \times Y \to \mathbb{R}$ be a $(\sigma(\mathfrak{A}, \mathfrak{B}), \mathfrak{Bor}_{\mathbb{R}})$-measurable function such that each

$$f_x = dP_x/d\nu.$$

In order to show that $f = d\Gamma/d(\alpha \times \nu)^*$, it is sufficient to show that f is integrable with respect to $\alpha \times \nu$ and that

$$\Gamma(A \times B) = \int_{A \times B} f\, d(\alpha \times \nu)$$

for all $A \times B \in \mathfrak{Semi}(\mathfrak{A}, \mathfrak{B})$, for then Γ extends uniquely to $\sigma(\mathfrak{A}, \mathfrak{B})$ by Proposition 8.8.

To show that f is integrable with respect to $\alpha \times \nu$, we will apply Tonelli's Theorem 5.19. We verify two things. First, each function

$$f_x : Y \to \mathbb{R} : y \mapsto f(x, y)$$

is integrable over Y with respect to ν for all $x \in X$, and this is because we have supposed $f_x = d\beta_x/d\nu$ for all $x \in X$. Second, the function

$$X \to \mathbb{R} : x \mapsto \int_Y f_x \, d\nu$$

defines an integrable function over X with respect to α since it is bounded and supposed to be $\big(\mathfrak{A}, \mathfrak{Bor}_{[0,1]}\big)$-measurable, and let us now go into more detail on this.

By hypothesis, $P^B(x) = P_x(B)$ for all $x \in X$, and for all $B \in \mathfrak{B}$, and in particular for $Y \in \mathfrak{B}$, and the assumption that $f_x = d\beta_x/d\nu$ for all $x \in X$ says that $P_x(B) = \int_B f_x \, d\nu$ for all $B \in \mathfrak{B}$. This means that

$$P^B(x) = P_x(B) = \int_B f_x \, d\nu,$$

which is bounded by 1 since each P_x is a probability measure, and which is $\big(\mathfrak{A}, \mathfrak{Bor}_{[0,1]}\big)$-measurable since we have supposed each $x \mapsto P^B(x)$ to be so. Consequently, the function

$$P^Y : X \to \mathbb{R} : x \mapsto \int_Y f_x \, d\nu$$

is integrable over X with respect to the measure α on $\mathfrak{A}$, and so certainly defines an integrable function over X with respect to α.

By Tonelli's Theorem 5.19 then, the function f is integrable with respect to $\alpha \times \nu$ and

$$\int_{A \times B} f \, d(\alpha \times \nu) = \int_A \int_B f_x \, d\nu \, d\alpha$$

for all $A \times B \in \mathfrak{Semi}(\mathfrak{A}, \mathfrak{B})$. By hypothesis,

$$\Gamma(A \times B) = \int_A P^B \, d\alpha,$$

for this is how we defined the joint Γ, and so by substitution for P^B:

$$\int_A P^B \, d\alpha = \int_A \int_B f_x \, d\nu \, d\alpha.$$

Therefore, by transitivity of equality,

$$\Gamma(A \times B) = \int_{A \times B} f \, d(\alpha \times \nu)$$

for all $A \times B \in \mathfrak{Semi}(\mathfrak{A}, \mathfrak{B})$, as required.

The fact that the function g equals $d\Gamma / d(\alpha \times \beta)^*$ follows by Proposition 5.26. ■

Proposition 9.4. *Let $\big(X \times Y, \sigma(\mathfrak{A}, \mathfrak{B}), \Gamma\big)$ be a probability space, and let ν be a σ-finite measure on $\mathfrak{B}$. Let $\Gamma \ll (\alpha \times \nu)^*$ on $\sigma(\mathfrak{A}, \mathfrak{B})$ with $f = d\Gamma / d(\alpha \times \nu)^*$. Define*

$$\alpha_y : \mathfrak{A} \to [0, 1] : A \mapsto \frac{\int_A f^y \, d\alpha}{\int_X f^y \, d\alpha}$$

for all $A \in \mathfrak{A}$, and for all $y \in Y$ such that f^y is integrable. If $\alpha \ll \lambda$ on $\mathfrak{A}$ with $h = d\alpha / d\lambda \in L_1(X, \mathfrak{A}, \lambda)$, then $\alpha_y \ll \lambda$ with

$$\frac{d\alpha_y}{d\lambda} = \frac{f^y h}{\int_X f^y h \, d\lambda},$$

or alternately

$$\alpha_y(A) = \frac{\int_A f^y h \, d\lambda}{\int_X f^y h \, d\lambda} \quad \textit{for all } A \in \mathfrak{A}.$$

Proof. Let $\alpha \ll \lambda$ on $\mathfrak{A}$ with $h = d\alpha/d\lambda \in L_1(X, \mathfrak{A}, \lambda)$. By definition,

$$\alpha_y(A) = \frac{\int_A f^y \, d\alpha}{\int_X f^y \, d\alpha},$$

which implies

$$\frac{d\alpha_y}{d\alpha} = \frac{f^y}{\int_X f^y \, d\alpha} \in L_1(X, \mathfrak{A}, \alpha).$$

It follows by Proposition 4.29 that

$$\frac{\int_A f^y \, d\alpha}{\int_X f^y \, d\alpha} = \frac{\int_A f^y h \, d\lambda}{\int_X f^y \, d\alpha},$$

and, again by Proposition 4.29, $\int_X f^y \, d\alpha = \int_X f^y h \, d\lambda$, so

$$\frac{\int_A f^y h \, d\lambda}{\int_X f^y \, d\alpha} = \frac{\int_A f^y h \, d\lambda}{\int_X f^y h \, d\lambda}.$$

By transitivity of equality,

$$\alpha_y(A) = \frac{\int_A f^y h \, d\lambda}{\int_X f^y h \, d\lambda},$$

as required. ■

9.3 Common Expression

You see the following type of combinations of mathematical symbols in colloquial texts:

$$p(\theta \mid y) = \frac{p(\theta)\, p(y \mid \theta)}{p(y)}. \tag{9.3}$$

Understand that expressions like these are only heuristic. Is there any way to interpret these symbols so that they make mathematical sense?

We suppose that the symbols indicate Bayes' theorem. Let us substitute θ for x:

$$\frac{d\alpha_y}{d\lambda}(\theta) = \frac{h(\theta)\, f(\theta, y)}{\displaystyle\int_\Omega h(\theta)\, f(\theta, y)\, d\lambda(\theta)}. \tag{9.4}$$

Now translate somewhat:

$\frac{d\alpha_y}{d\lambda}(\theta)$	$p(\theta \mid y)$	posterior density
$h(\theta) = \frac{d\alpha}{d\lambda}(\theta)$	$p(\theta)$	prior density
$f(\theta, y)$	$p(y \mid \theta)$	likelihood, sort of
$\int_\Omega h(\theta)\, f(\theta, y)\, d\lambda(\theta)$	$p(y)$	various labels

In order for the function $f(\theta, y)$ to make sense in Bayes' theorem, as in equation (9.4), it should be written as $f^y(\theta)$, since y is fixed, and $f^y : \Theta \to \mathbb{R}$. But the function f^y need not be a density.

In order for the function $f(\theta, y)$ to make sense as a density then, it should be written as $f_\theta(y)$; this is because θ is fixed, and as stated in equation 9.1:

$$f_\theta = dP_\theta / d\nu,$$

which is a Radon-Nikodym derivative. But f_θ is a map $Y \to \mathbb{R}$, and so f_θ makes no sense in Bayes' theorem.

The question then stands: In a heuristic expression such as (9.3), do the authors somehow intend that the symbols $p(y \mid \theta)$ be interpreted as a density?

Heuristic expressions can introduce some amusing problems.

10 Bayesian Statistical Models

10.1 Outline

Here is a partial list of steps that will be applied to each of the examples that follow:

- Establish an indexed, or parametrized, family $\{P_x : x \in X\}$ of sampling probabilities on the sample space σ-algebra $\mathfrak{B}$. Hopefully, each P_x is absolutely continuous with respect to a σ-finite reference measure on $\mathfrak{B}$.
- Verify that some function $f : X \times Y \to \mathbb{R}$ has f_x as the density function for P_x with respect to the reference measure.
- Define a prior probability measure on the parameter space σ-algebra $\mathfrak{A}$. Say that h is a density function for the prior with respect to a σ-finite reference measure on $\mathfrak{A}$, temporarily call this measure ν.
- Then understand that the map

$$X \to \mathbb{R} : x \mapsto \frac{f(x,y)h(x)}{\displaystyle\int_X f(x,y)h(x)\,d\nu(x)} = \frac{f^y(x)h(x)}{\displaystyle\int_X f^y(x)h(x)\,d\nu(x)}$$

  is a density function for the posterior probability measure of the parameter given the data, with respect to ν on $\mathfrak{A}$.

10.2 Globe Tossing Model

This example of an experiment is highly paraphrased from Richard McElreath's *Statistical Rethinking, Second Edition* [6]. The experiment is tossing a globe N times into the air, catching it, and wherever an index finger happens to touch the surface of the globe, record *water* should this surface corresponds to water, and record *land* should it correspond to land. This N-tuple is the sample. Deduce from this experiment the probability distribution of the proportion p of the surface covered with water. This proportion p is the parameter. Quoting from *Statistical Rethinking* [6], page 78, verbatim:

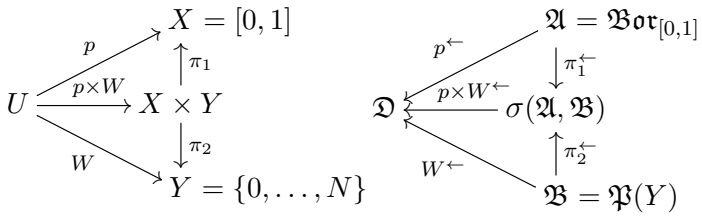
> **4.2.1. Re-describing the globe tossing model.** It's good to work with examples. Recall the proportion of water problem from previous chapters. The model in that case was always:
>
> $$W \sim \text{Binomial}(N, p)$$
> $$p \sim \text{Uniform}(0, 1)$$
>
> where W was the observed count of water, N was the total number of tosses, and p was the proportion of water on the globe. Read the above statement as:
>
> > The count W is distributed binomially with sample size N and probability p.
> > The prior for p is assumed to be uniform between zero and one.

And in a display following this, find:

$$\Pr(p|w, n) = \frac{\text{Binomial}(w|n, p)\,\text{Uniform}(p|0, 1)}{\int \text{Binomial}(w|n, p)\,\text{Uniform}(p|0, 1)\,dp}$$

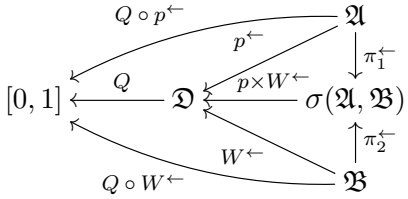
Let U denote the domain of the measurable functions p and W,

as illustrated:



so that the underlying space U is the Cartesian product of the interval $[0, 1]$, representing all possible proportions of water to land, with the collection of all N-tuples of *water*'s and *land*'s. The function W is supposed to count the number of *water*'s in an N-tuple of *water*'s and *land*'s. Let $\mathfrak{D}$ denote a σ-algebra of subsets of U, to be qualified momentarily. Then the parameter space X and sample space Y are the respective codomains, where $X = [0, 1]$ with σ-algebra $\mathfrak{A} = \mathfrak{Bor}_{[0,1]}$, and Y is the set of integers $\{0, \ldots, N\}$ with σ-algebra $\mathfrak{B} = \mathfrak{P}(Y)$.

We suppose $(U, \mathfrak{D}, Q)$ is a probability space such that the following diagram commutes:



We can take $\mathfrak{D}$ to be the σ-algebra generated by the product of $[0, 1]$ with the collection of N-tuples of *water*'s and *land*'s. The probability distributions can be specified, although the specification of the

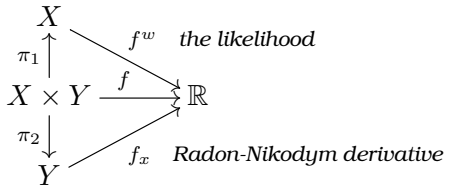
measure Q appears induced, and will not be stressed here:

$$W \sim \text{Binomial}(N, x) = Q \circ W^{\leftarrow}$$
$$p \sim \text{Uniform}(0, 1) = Q \circ p^{\leftarrow}$$

Define $f : X \times Y \to \mathbb{R}$ by

$$\begin{aligned} f(x, w) &= \frac{N!}{w!(N-w)!} x^w (1-x)^{N-w} \\ &=: \text{Binomial}(w|N, x) \end{aligned}$$

so the following diagram commutes:



Let κ denote counting measure on the sample space σ-algebra $\mathfrak{B}$, and define a family $\{P_x : x \in X\}$ of probability measures on $\mathfrak{B}$ by

$$P_x(B) = \int_{w \in B} f_x(w)\, d\kappa(w) = \sum_{w \in B} f_x(w) \quad B \in \mathfrak{B}.$$

Let us verify one of the requirements that a measure must satisfy in order to be a probability measure; namely that $P_x(Y) = 1$ for any $x \in X$. This verification is based on the binomial theorem:

$$(a+b)^N = \sum_{w=0}^{N} \frac{N!}{w!(N-w)!} a^w b^{N-w},$$

taking $a = x$ and $b = 1 - x$.

By definition, each of the sampling probabilities P_x is absolutely continuous with respect to counting measure κ, with $f_x = dP_x/d\kappa$. That is, the function f_x represents the Radon-Nikodym derivative of P_x with respect to κ.

Let α be the probability measure on the parameter space σ-algebra $\mathfrak{A} = \mathfrak{Bor}_{[0,1]}$ represented by the density

$$h(x) = \text{Uniform}(x|0,1)$$

with respect to Lebesgue measure ν restricted to $\mathfrak{Bor}_{[0,1]}$. In this case, the function h agrees with the constant function 1 on $[0,1]$, and the prior α also happens to agree with Lesbesgue measure ν on $\mathfrak{Bor}[0,1]$. Namely,

$$\alpha(A) = \int_A h(x)\, d\nu(x) \quad \text{for all } A \in \mathfrak{Bor}_{[0,1]}.$$

According to Bayes' theorem, the Radon-Nikodym derivative of the posterior probability α_w with respect to ν has a representative, denoted here by $\Pr(x|w,N) \in \mathscr{L}_1(X, \mathfrak{A}, \nu)$, which may be called the *posterior density:*

$$\Pr(x|w,N) \coloneqq \frac{f^w(x)h(x)}{\int_X f^w(x)h(x)\, d\nu(x)} \in \frac{d\alpha_w}{d\nu}.$$

Finally, this can also be written as:

$$\Pr(x|w,N) = \frac{\overbrace{\text{Binomial}(w|N,x)}^{f^w(x)} \overbrace{\text{Uniform}(x|0,1)}^{h(x)}}{\int_{[0,1]} \text{Binomial}(w|N,x) \underbrace{\text{Uniform}(x|0,1)\, dx}_{h(x)\, d\nu(x)}},$$

which is not difficult to see as a slight translation of

$$\Pr(p|w,n) = \frac{\text{Binomial}(w|n,p)\, \text{Uniform}(p|0,1)}{\int \text{Binomial}(w|n,p)\, \text{Uniform}(p|0,1)\, dp}.$$

10.3 Gaussian Model of Height

Another example of an experiment with the joint absolutely continuous with respect to the product of the prior and predictive probability distributions, paraphrased from *Statistical Rethinking, Second Edition* [6]. The experiment is modelling a collection $\{h_i\}$ of heights. Quoting, beginning bottom page 78:

> For the moment, we want a single measurement variable to model as a Gaussian distribution. There will be two parameters describing the distribution's shape, the mean μ and the standard deviation σ.

and continuing on page 82:

> In most cases, priors are specified independently for each parameter, which amounts to assuming $\Pr(\mu, \sigma) = \Pr(\mu)\Pr(\sigma)$. Then we can write:

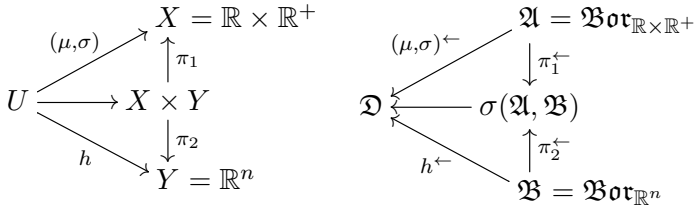
$$\begin{aligned} h_i &\sim \text{Normal}(\mu, \sigma) && \text{[likelihood]} \\ \mu &\sim \text{Normal}(178, 20) && [\mu \text{ prior}] \\ \sigma &\sim \text{Uniform}(0, 50) && [\sigma \text{ prior}] \end{aligned}$$

and in a display following, find:

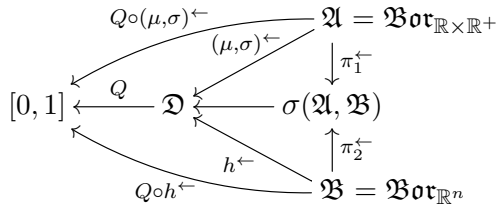
$$\begin{aligned} &\Pr(\mu, \sigma | h) = \\ &= \frac{\prod_i \text{Normal}(h_i|\mu, \sigma)\, \text{Normal}(\mu|178, 20)\, \text{Uniform}(\sigma|0, 50)}{\int \int \prod_i \text{Normal}(h_i|\mu, \sigma)\, \text{Normal}(\mu|178, 20)\, \text{Uniform}(\sigma|0, 50)\, d\mu\, d\sigma} \end{aligned}$$

Let U denote the domain of the functions (μ, σ) and height h, as

illustrated:



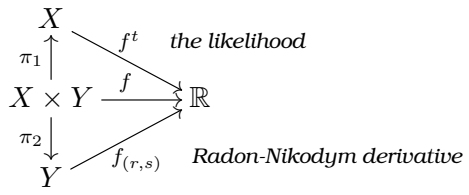
so that the underlying space U is the Cartesian product of possible means and standard deviations with n-tuples of heights. We suppose $(U, \mathfrak{D}, Q)$ is a probability space such that the following diagram commutes:



Define $f : X \times Y \to \mathbb{R}$ by

$$f\big((r,s),t\big) = \prod_i \frac{1}{s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t_i - r}{s}\right)^2\right\}$$

so the following diagram commutes:



Let λ_n denote Lebesgue measure on the sample space σ-algebra $\mathfrak{B} = \mathfrak{Bor}_{\mathbb{R}^n}$. Define a family $\{P_{(r,s)} : (r,s) \in X\}$ of probability measures on $\mathfrak{B}$ by defining $P_{(r,s)}$ on the rectangles $\prod_i [a_i, b_i]$ of $\mathfrak{B}$:

$$P_{(r,s)}\left(\prod_i [a_i, b_i]\right) = \prod_i \text{Normal}(r,s)\big([a_i, b_i]\big)$$

Let us verify one of the requirements that a measure must satisfy in order to be a probability measure; namely that $P_{(r,s)}(Y) = 1$ for any $(r,s) \in X$. This verification is based on the fact that the functions

$$p : \mathbb{R} \to \mathbb{R}^+ : t \mapsto \frac{1}{s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-r}{s}\right)^2\right\}$$

are well-known to be a probability density functions.

Each of the sampling probabilities $P_{(r,s)}$ is absolutely continuous with respect to Lebesgue measure λ_n on $\mathfrak{Bor}_{\mathbb{R}^n}$, and the function $f_{(r,s)}$ represents the Radon-Nikodym derivative of $P_{(r,s)}$ with respect to λ_n.

Define a probability measure ρ on the rectangles $B \times C$ of the parameter space σ-algebra $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}\times\mathbb{R}^+}$ by

$$\rho(B \times C) = \text{Normal}(178, 20)(B) \cdot \text{Uniform}(0, 50)(C).$$

See that the prior probability distribution ρ is absolutely continuous with respect to a different Lebesgue measure, call it λ_2, on $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}\times\mathbb{R}^+}$, and let $h \in \mathscr{L}_1(X, \mathfrak{A}, \lambda_2)$ be defined by

$$h : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} : (r,s) \mapsto \text{Normal}(r|178, 20) \cdot \text{Uniform}(s|0, 50)$$

so that the function h represents the Radon-Nikodym derivative $d\rho/d\lambda_2$. According to Bayes' theorem, the Radon-Nikodym deriva-

tive of the posterior probability ρ_t with respect to λ_2 has a representative, denoted here by $\Pr(r, s|t) \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$, where

$$\Pr(r, s|t) = \frac{f^t(r, s)\, h(r, s)}{\int_X f^t(r, s)\, h(r, s)\, d\lambda_2(r, s)}.$$

Finally, write

$$\begin{aligned} f^t(r, s) &= \prod_i \frac{1}{s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t_i - r}{s}\right)^2\right\} \\ &= \prod_i \text{Normal}(t_i|r, s) \end{aligned}$$

so that

$$\Pr(r, s|t) = \frac{\overbrace{\prod_i \text{Normal}(t_i|r, s)}^{f^t(r,s)} \overbrace{\text{Normal}(r|178, 20)\, \text{Uniform}(s|0, 50)}^{h(r,s)}}{\int_{\mathbb{R}\times\mathbb{R}^+} \prod_i \text{Normal}(t_i|r, s)\, \underbrace{\text{Normal}(r|178, 20)\, \text{Uniform}(s|0, 50)\, d\lambda_2(r, s)}_{h(r,s)\, d\lambda_2(r,s)}}$$

which is our slight translation of

$$\begin{aligned} &\Pr(\mu, \sigma|h) = \\ &= \frac{\prod_i \text{Normal}(h_i|\mu, \sigma)\, \text{Normal}(\mu|178, 20)\, \text{Uniform}(\sigma|0, 50)}{\int\int \prod_i \text{Normal}(h_i|\mu, \sigma)\, \text{Normal}(\mu|178, 20)\, \text{Uniform}(\sigma|0, 50)\, d\mu\, d\sigma}. \end{aligned}$$

10.4 The Linear Model

The linear model we will use here can be seen as an extension of the Gaussian model in Example 10.3 by including weight, assuming a relation between height and weight. The word *relation* is used in the mathematical sense; it is a subset of a Cartesian product. The model is given in this quote from *Statistical Rethinking*, page 93:

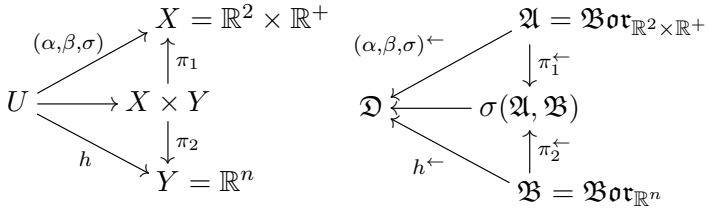
> Now how do we get weight into a Gaussian model of height? Let x be the name for the column of weight measurements, `d2$weight`. Let the average of the x values be $\bar{x}$, "ex bar". Now we have a predictor variable x, which is a list of measures of the same length as h. To get weight into the model, we define the mean μ as a function of the values in x. This is what it looks like, with explanation to follow:

$$
\begin{aligned}
h_i &\sim \text{Normal}(\mu_i, \sigma) && \text{[likelihood]} \\
\mu_i &= \alpha + \beta(x_i - \bar{x}) && \text{[linear model]} \\
\alpha &\sim \text{Normal}(178, 20) && [\alpha \text{ prior}] \\
\beta &\sim \text{Normal}(0, 10) && [\beta \text{ prior}] \\
\sigma &\sim \text{Uniform}(0, 50) && [\sigma \text{ prior}]
\end{aligned}
$$

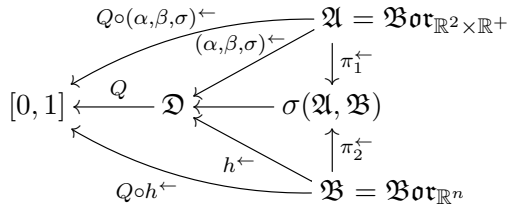
See that this model includes no probability distribution for the weight measurements x_i, but the measurable functions h_i, α, β, and σ do have probability distributions included in the model. We consider the x_i as known values.

Let U denote the domain of the functions (α, β, σ) and h, as

illustrated:



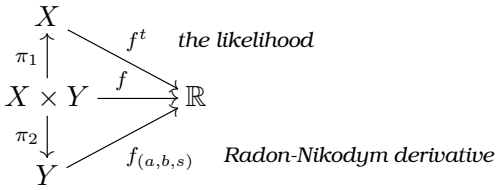
We suppose $(U, \mathfrak{D}, Q)$ is a probability space such that the following diagram commutes:



Define $f : X \times Y \to \mathbb{R}$ by

$$\begin{aligned} f\big((a,b,s),t\big) &= \prod_i \frac{1}{s\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{t_i - \big(a \cdot 1 + b(x_i - \bar{x})\big)}{s} \right)^2 \right\} \\ &= \prod_i \frac{1}{s\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{t_i - \mu_i}{s} \right)^2 \right\} \\ &= \prod_i \text{Normal}(t_i | \mu_i, s), \end{aligned}$$

where $\mu_i = a \cdot 1 + b(x_i - \bar{x})$, so the following diagram commutes:



Define a family $\{P_{(a,b,s)} : (a, b, s) \in X\}$ of measures on the sample space σ-algebra $\mathfrak{B} = \mathfrak{Bor}_{\mathbb{R}^n}$ in terms of the indefinite integrals of the collection $\{f_{(a,b,s)} : (a, b, s) \in X\}$, integrated with respect to Lebesgue measure, call it λ_n, on $\mathfrak{Bor}_{\mathbb{R}^n}$;

$$P_{(a,b,s)}(B) = \int_B f_{(a,b,s)}\, d\lambda_n \quad \text{for all } B \in \mathfrak{Bor}_{\mathbb{R}^n}.$$

Let us verify a requirement that a measure must satisfy in order to be a probability measure; namely that $P_{(a,b,s)}(Y) = 1$ for any $(a, b, s) \in X$. This is based on the fact that the functions

$$p : \mathbb{R} \to \mathbb{R}^+ : t \mapsto \frac{1}{s\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-r}{s}\right)^2\right\}$$

are well-known to be a probability density functions.

For the purpose of comparison, see that the sum $a_1^2 + a_2^2 + a_3^2$ can be written as $A^\intercal A$ with $A^\intercal = [a_1\ a_2\ a_3]$. Were you to then write $a_i = y_i - (1 \cdot \beta_0 + x_i \cdot \beta_1)$, with $\beta = (\beta_0, \beta_1)$, and with X and Y now functions or vectors or variables or whatever, you would end up with something like:

$$f(\beta, \sigma^2, Y) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}[Y - X\beta]^\intercal [Y - X\beta]\right\},$$

which you could compare with equation (3.1) on page 51 of Marin & Robert, *Bayesian Core*, which says the likelihood of the *ordinary normal linear model* is

$$l(\beta, \sigma^2 | y, X) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^\intercal (y - X\beta)\right\}.$$

The translation between their point of view and ours is not exact. For one thing, our a and b would be part of their β. For another, if the particular data included the y-values at known X values, say when X is equal to X_0, then our f would relate to their l by something like:

$$f(\beta, \sigma^2, y) = l(\beta, \sigma^2 | y, X)\Big|_{X=X_0}$$

Define a probability measure ρ on the rectangles $A \times B \times C$ of the parameter space σ-algebra $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^2 \times \mathbb{R}^+}$ by

$$\begin{aligned}&\rho(A \times B \times C) = \\ &= \text{Normal}(178, 20)(A) \cdot \text{Normal}(0, 10)(B) \cdot \text{Uniform}(0, 50)(C).\end{aligned}$$

See that the prior probability distribution ρ is absolutely continuous with respect to a different Lebesgue measure, call it λ, on $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^2 \times \mathbb{R}^+}$, and let $h : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R}$ in $\mathscr{L}_1(X, \mathfrak{A}, \lambda)$ be defined by

$$h(a, b, s) = \text{Normal}(a|178, 20) \cdot \text{Normal}(b|0, 20) \cdot \text{Uniform}(s|0, 50)$$

so that h represents the Radon-Nikodym derivative $d\rho/d\lambda$. According to Bayes' theorem, the Radon-Nikodym derivative of the posterior probability ρ_t with respect to λ has a representative, denoted here by $\Pr(a, b, s \,|\, t) \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$, where

$$\Pr(a, b, s \,|\, t) = \frac{f^t(a, b, s)\, h(a, b, s)}{\int_X f^t(a, b, s)\, h(a, b, s)\, d\lambda(a, b, s)},$$

with

$$\mu_i = a \cdot 1 + b(x_i - \bar{x})$$

and with

$$f^t(a, b, s) = \prod_i \frac{1}{s\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{t_i - \mu_i}{s} \right)^2 \right\}.$$

10.5 Logistic Regression

The logit function is defined by

$$\text{logit} : (0,1) \to \mathbb{R} : p \mapsto \ln\left(\frac{p}{1-p}\right),$$

and the logistic function is defined by

$$\text{logistic} : \mathbb{R} \to (0,1) : a \mapsto \frac{\exp(a)}{1+\exp(a)}.$$

The logit function is the inverse of the logistic function, that is,

$$\text{logit}(p) = a \quad \Leftrightarrow \quad p = \text{logistic}(a),$$

which says

$$\ln\left(\frac{p}{1-p}\right) = a \quad \Leftrightarrow \quad p = \frac{\exp(a)}{1+\exp(a)}.$$

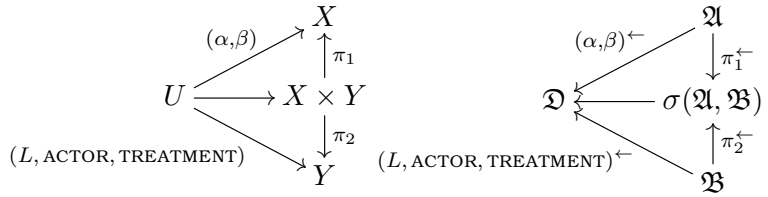
The expression $p/(1-p)$ can be seen as the odds ratio.

The logistic regression model we will look at here, quoted from the tentative model displayed in *Statistical Rethinking,* page 327, is:

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{ACTOR}[i]} + \beta_{\text{TREATMENT}[i]} \\
\alpha_j &\sim \text{ to be determined} \\
\beta_k &\sim \text{ to be determined}
\end{aligned}
$$

The "$\alpha_j \sim$ to be determined" will later be $\alpha_j \sim \text{Normal}(0, 1.5)$, and "$\beta_k \sim$ to be determined" will later be $\beta_k \sim \text{Normal}(0, 0.5)$.

Let $(\alpha, \beta) : U \to X$ and $(L, \text{ACTOR}, \text{TREATMENT}) : U \to Y$ be maps such that the following diagrams commute:



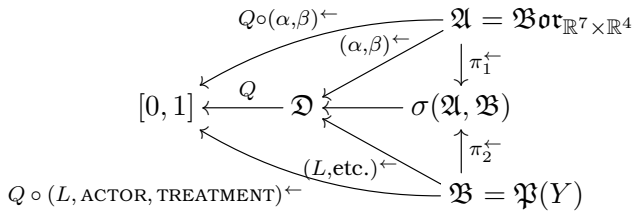
with

$$X = \mathbb{R}^7 \times \mathbb{R}^4$$

and

$$Y = \{0, 1\}^n \times \{1, \ldots, 7\}^n \times \{1, 2, 3, 4\}^n,$$

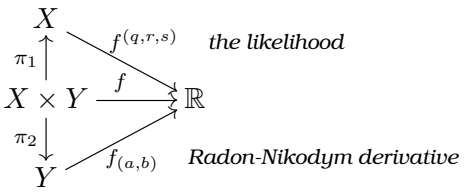
where $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^7 \times \mathbb{R}^4}$ and $\mathfrak{B} = \mathfrak{P}(Y)$. We are supposing $(U, \mathfrak{D}, Q)$ is a probability space such that the following diagram commutes:



Define $f : X \times Y \to \mathbb{R}$ by

$$\begin{aligned}\big((a,b),(q,r,s)\big) &\mapsto \prod_i \big(\text{logistic}(a_{r_i} + b_{s_i})\big)^{q_i} \big(1 - \text{logistic}(a_{r_i} + b_{s_i})\big)^{1-q_i} \\ &= \prod_i \text{Binomial}(q_i|1,p_i), \qquad \text{with } p_i = \text{logistic}(a_{r_i} + b_{s_i})\end{aligned}$$

so the following diagram commutes



Define a family $\{P_{(a,b)} : (a,b) \in X\}$ of measures on the sample space σ-algebra $\mathfrak{B} = \mathfrak{P}(Y)$ in terms of the indefinite integrals of the collection $\{f_{(a,b)} : (a,b) \in X\}$, integrated with respect to counting measure on $\mathfrak{P}(Y)$, calling counting measure δ here;

$$\begin{aligned}P_{(a,b)}(B) &= \int_B f_{(a,b)}\, d\delta \\ &= \sum_{(q,r,s)\in B} f_{(a,b)}(q,r,s) \\ &= \sum_{(q,r,s)\in B} \prod_i \text{Binomial}(q_i|1,p_i),\end{aligned}$$

with $p_i = \text{logistic}(a_{r_i} + b_{s_i})$. These measures are probability measures because they are defined in terms of the density

$$\prod_i \text{Binomial}(q_i|1,p_i)$$

of a probability measure. Roughly.

Define a probability measure ρ on the rectangles $A \times B$ of the parameter space σ-algebra $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^7 \times \mathbb{R}^4}$ by

$$\rho(A \times B) = \text{Normal}(0, 1.5)(A) \cdot \text{Normal}(0, 0.5)(B).$$

The prior probability distribution ρ is absolutely continuous with respect to a Lebesgue measure, call it λ, on $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^7 \times \mathbb{R}^4}$. Let $h \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$ be defined by

$$h : \mathbb{R}^7 \times \mathbb{R}^4 \to \mathbb{R} : (a, b) \mapsto \text{Normal}(a|0, 1.5) \cdot \text{Normal}(b|0, 0.5).$$

so that h represents the Radon-Nikodym derivative $d\rho/d\lambda$. According to Bayes' theorem, the Radon-Nikodym derivative of the posterior probability $\rho_{(q,r,s)}$ with respect to λ has a representative, denoted here by $\Pr(a, b \,|\, q, r, s) \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$, where

$$\Pr(a, b \,|\, q, r, s) = \frac{f^{(q,r,s)}(a, b)\, h(a, b)}{\int_X f^{(q,r,s)}(a, b)\, h(a, b)\, d\lambda(a, b)}.$$

10.6 Toy Logistic Regression

Let us apply logistic regression to a very simple model.

> The initial purpose of this "toy" approach was to set up a template for dropping in a much larger model. The larger model was to be the Criteo display ad Kaggle challenge.
>
> The idea behind the Kaggle competition would be to predict whether or not an online advertisement will be mouse-clicked, given a data set having 13 integer-valued features and 26 categorical features. There is a set of *training* data, meaning we are given the feature values and also told whether the corresponding ad was clicked on. Then there is a *test* set of

data, meaning we are given the feature values, but need to predict whether the corresponding ad would be clicked on.

Both training and test data sets have missing values. The training data set has around 45 million rows, or observations, with 90% of those rows having some missing value. Around 40 million missing values, at least.

Another difficulty: The test set data also has many missing values. I am not sure right now how to make predictions using data sets loaded with missing values. There are at least two ways to approach this:

1. The MEASURE THEORETIC METHOD: Do a full-on, measure-theoretic analysis of Bayesian prediction with missing values in the data set. Right down to the σ-algebras, which does not seem to have appeared yet in the literature. In other words, I gotta do my homework.
2. The HELL'S BELLS METHOD: Statistics is already guesswork. Just use any of the existing missing values methods, and call it good. Why even think about the σ-algebras?

Probably need to do both for comparison. Did I mention this was a works-in-progress? Back to toy logistic regression.

Start with something simple. Suppose some FEATURE is a list of real values, and a function CLICK represents success or failure, depending somewhat on the FEATURE values. We think of CLICK as representing whether a computer mouse was clicked upon. Write the probability model as:

$$\begin{aligned} \text{CLICK}_i &\sim \text{Binomial}(1, p_i) \\ \text{logit}(p_i) &= \alpha + \beta * \text{FEATURE}_i \\ \alpha &\sim \text{Normal}(0, 1.6) \\ \beta &\sim \text{Normal}(0, 1.6) \end{aligned}$$

From a mathematician's point of view, we want to expose all the math so we can clearly see what is going on. Let U denote an underlying probability space. Define the parameter space $X \coloneqq \mathbb{R}^2$, and let

$$(\alpha, \beta) : U \to X : u \mapsto (a, b).$$

Define the sample space $Y \coloneqq \{0,1\}^3$, and let

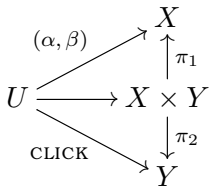
$$\text{CLICK} : U \to Y : u \mapsto q.$$

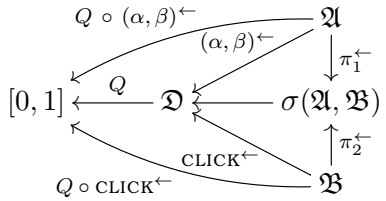
Also let

$$\text{FEATURE} : U \to \mathbb{R}^3 : u \mapsto r,$$

but we treat the FEATURE values as known, without error, meaning FEATURE does not map to a probability distribution in our probability model.

We are supposing that $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^2}$ and $\mathfrak{B} = \mathfrak{P}(\{0,1\}^3)$, and $(U, \mathfrak{D}, Q)$ is a probability space such that the following diagrams commute:

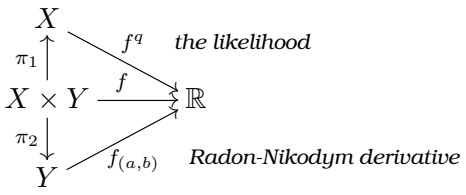




Treating the FEATURE values as known, meaning we consider the r_i as constants, define $f : X \times Y \to \mathbb{R}$ by

$$\begin{aligned}\big((a,b),q\big) &\mapsto \prod_{i=1}^{3} \big(\text{logistic}(a + b \cdot r_i)\big)^{q_i} \big(1 - \text{logistic}(a + b \cdot r_i)\big)^{1-q_i} \\ &= \prod_{i=1}^{3} \text{Binomial}(q_i | 1, p_i), \qquad \text{with } p_i = \text{logistic}(a + b \cdot r_i)\end{aligned}$$

and illustrating:



Define a family $\{P_{(a,b)} : (a,b) \in X\}$ of probability measures on the sample space σ-algebra $\mathfrak{B}$ in terms of the indefinite integrals of the collection $\{f_{(a,b)} : (a,b) \in X\}$, integrated with respect to counting measure κ on $\mathfrak{B}$:

$$\begin{aligned} P_{(a,b)}(B) &= \int_B f_{(a,b)}(q)\, d\kappa \\ &= \sum_{(a,b)\in B} f_{(a,b)}(q) \\ &= \sum_{(a,b)\in B} \prod_{i=1}^{3} \text{Binomial}(q_i|1,p_i), \qquad \text{with } p_i = \text{logistic}(a + b \cdot r_i) \end{aligned}$$

Although it may be difficult to tell, each of these probability measures are really just an example of a multivariate Bernoulli distribution, a distribution that must have been known since time immemorial. The sum of a such a probability distribution over a sample space, in this case Y, is known to be 1, but let us verify this. That is, we will verify that

$$\sum_{y\in Y} \prod_{i=1}^{3} p_i^{y_i} (1-p_i)^{(1-y_i)} = 1$$

y_1	y_2	y_3	$\prod p_i^{y_i}(1-p_i)^{(1-y_i)}$
0	0	0	$(1-p_1)(1-p_2)(1-p_3)$
0	0	1	$(1-p_1)(1-p_2)p_3$
0	1	0	$(1-p_1)p_2(1-p_3)$
0	1	1	$(1-p_1)p_2p_3$
1	0	0	$p_1(1-p_2)(1-p_3)$
1	0	1	$p_1(1-p_2)p_3$
1	1	0	$p_1p_2(1-p_3)$
1	1	1	$p_1p_2p_3$

Table 10.1: Verify the sum of products is 1.

by listing the products in Table 10.1, and then computing their sum.

The sum of the first two lines from Table 10.1, factoring out the $(1-p_1)(1-p_2)$, is

$$(1-p_1)(1-p_2)\big((1-p_3)+p_3\big) = (1-p_1)(1-p_2).$$

The sum of the third and fourth lines from Table 10.1, factoring out the $(1-p_1)p_2$, is

$$(1-p_1)p_2\big((1-p_3)+p_3\big) = (1-p_1)p_2.$$

So the sum of the first four lines from the table is

$$\begin{aligned}(1-p_1)(1-p_2)+(1-p_1)p_2 &= (1-p_1)\big((1-p_2)+p_2\big)\\ &= 1-p_1.\end{aligned}$$

Similarly, the sum of the last four lines from the table is p_1, and so the sum of all eight lines from the table is $1-p_1+p_1=1$, as claimed. Yes, I know this is incredibly simple algebra, and you could have done this in your head.

Define the prior probability measure ρ on the rectangles $A \times B$ of the parameter space σ-algebra $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^2}$ by

$$\rho(A \times B) = \text{Normal}(0, 1.6)(A) \cdot \text{Normal}(0, 1.6)(B).$$

The probability distribution ρ is absolutely continuous with respect to a Lebesgue measure, call it λ, on $\mathfrak{A} = \mathfrak{Bor}_{\mathbb{R}^2}$. Let $h \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$ be defined by

$$h : \mathbb{R}^2 \to \mathbb{R} : (a, b) \mapsto \text{Normal}(a|0, 1.6) \cdot \text{Normal}(b|0, 1.6).$$

so that h represents the Radon-Nikodym derivative $d\rho/d\lambda$. According to Bayes' theorem, the Radon-Nikodym derivative of the posterior probability ρ_q with respect to λ has a representative, denoted here by $\Pr(a, b \,|\, q) \in \mathscr{L}_1(X, \mathfrak{A}, \lambda)$, where $p_i = \text{logistic}(a + b \cdot r_i)$ and

$$\begin{aligned}
&\Pr(a, b \,|\, q) = \\
&= \frac{f^q(a, b)\, h(a, b)}{\int_X f^q(a, b)\, h(a, b)\, d\lambda(a, b)} \\
&= \frac{\prod_i \text{Binomial}(q_i|1, p_i)\, \text{Normal}(a|0, 1.6)\, \text{Normal}(b|0, 1.6)}{\int_{\mathbb{R}^2} \prod_i \text{Binomial}(q_i|1, p_i)\, \text{Normal}(a|0, 1.6)\, \text{Normal}(b|0, 1.6)\, d\lambda(a, b)}.
\end{aligned}$$

Just to see how some corresponding R code might look:

R code 10.1

```
#logistic regression on custom data
#set up your data frame first, make three points:
```

```
x <- c(1,2,3)
y <- c(0,0,1)
d <- data.frame(list(feature=x, click=y))

library(rethinking)

m1 <- quap(
 alist(
  click ~ dbinom(1,p),
  logit(p) <- a + b*feature,
  a ~ dnorm(0, 1.6),
  b ~ dnorm(0, 1.6)
 ), data=d
)
```

10.7 A/B Testing: The Code for Facial Identity in the Primate Brain

In the paper, *The Code for Facial Identity in the Primate Brain* [1], the authors use recordings from 205 neurons to construct a model. The model describes how facial identity might be represented in the primate brain. In this case, macaques. It is based upon the idea that when the macaque sees a face, the response of those neurons is a function of a linear combination of facial features, those features comprising a basis, or axis.

Are these neurons tuned to respond to facial images projected onto various axes, or are they tuned to respond to "exemplar" faces? This issue is addressed in the paper, and should give us a good chance to do some Bayesian A/B testing. That is, to compare the

axis model with the exemplar model. It may also give us a chance to compare the paper's frequentist tests with Bayesian methods.

To be continued.

11 Expectation Operators as Projections

AN ORTHOGONAL projection on a Hilbert space is a linear idempotent whose kernel is orthogonal to its range. There is a one-to-one correspondence between the closed subspaces of a Hilbert space and its orthogonal projections. Let $\mathfrak{H}$ be a Hilbert space, and let $\mathfrak{M}$ be a closed subspace of $\mathfrak{H}$. To each $x \in \mathfrak{H}$ there corresponds a unique element of $\mathfrak{M}$, denoted here by $\mathfrak{Proj}(x)$, such that $x - \mathfrak{Proj}(x) \perp \mathfrak{M}$. The map $\mathfrak{H} \to \mathfrak{H}$ defined by $x \mapsto \mathfrak{Proj}(x)$ is called the orthogonal projection of $\mathfrak{H}$ onto $\mathfrak{M}$ along $\mathfrak{M}^{\perp}$, or simply the *orthogonal projection of* $\mathfrak{H}$ *onto* $\mathfrak{M}$.

Orthogonal projections are contractive in the sense that any nonzero orthogonal projection on a Hilbert space has norm equal to 1. For x and y in $\mathfrak{H}$, we let $\langle x, y \rangle$ denote their inner product.

Proposition 11.1. *Let $\mathfrak{H}$ be a Hilbert space, and let $\mathfrak{M}$ be a closed subspace of $\mathfrak{H}$, where $\mathfrak{Proj} : \mathfrak{H} \to \mathfrak{H}$ is the orthogonal projection of $\mathfrak{H}$ onto $\mathfrak{M}$. Fix $x \in \mathfrak{H}$. Then $\mathfrak{Proj}(x)$ is that element of $\mathfrak{M}$ which is closest to x, meaning that*

$$\|x - \mathfrak{Proj}(x)\| < \|x - y\|$$

for all y in the subspace $\mathfrak{M}$ such that $\mathfrak{Proj}(x) \neq y$.

Proposition 11.2. *Let $\{e_1, e_2, \ldots, e_n\}$ be an orthonormal set in a Hilbert space $\mathfrak{H}$, and let $\mathfrak{Proj} : \mathfrak{H} \to \mathfrak{H}$ be the orthogonal projection of $\mathfrak{H}$*

onto the $\mathfrak{Span}\{e_1, e_2, \ldots, e_n\}$*. Fix* $x \in \mathfrak{H}$*. Then* $\mathfrak{Proj}(x) = \sum_{i=1}^{n} \langle x, e_i \rangle e_i$*. That is,*

$$\left\| x - \sum_{i=1}^{n} a_i e_i \right\| < \|x - y\|$$

for all y *in the* $\mathfrak{Span}\{e_1, e_2, \ldots, e_n\}$ *such that* $\mathfrak{Proj}(x) \neq y$ *if and only if each scalar* a_i *equals the Fourier coefficient* $\langle x, e_i \rangle$ *of* x*.*

11.1 Expectation as Projection

In order to complete this section, you should look at the case in which the σ-subalgebra is generated by a finite partition. After all, this is always the case you will be using when doing any computer modeling.

It means wheeling in a decent exposition of the finite case. Use Cheng-Shang Chang's, *Understanding Conditional Expectation via Vector Projection.* Then decide how far you want to go with this. Can you do better than *A Deep Dive Into How R Fits a Linear Model* from `http://madrury.github.io/`?

Discussion 11.3. Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The conditional expectation operator $\mathscr{E}_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1(X, \mathfrak{A}, \mu)$ is idempotent; if $\overline{f}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$, then $\mathscr{E}_{\mathfrak{B}} \overline{f}^{\mu} \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, and so by Proposition 7.29,

$$\mathscr{E}_{\mathfrak{B}} \overline{f}^{\mu} = \mathscr{E}_{\mathfrak{B}} (\mathscr{E}_{\mathfrak{B}} \overline{f}^{\mu}).$$

That is, $\mathscr{E}_{\mathfrak{B}} = \mathscr{E}_{\mathfrak{B}} \circ \mathscr{E}_{\mathfrak{B}}$ on $L_1(X, \mathfrak{A}, \mu)$.

The inclusion in Discussion 11.4 is really an embedding; see Discussion 4.16. Change your subsets to embeddings.

Discussion 11.4. Let us see how the operator $\mathscr{E}_{\mathfrak{B}}$ restricts to $L_2(X, \mathfrak{A}, \mu)$. The subspace $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ of the space $L_2(X, \mathfrak{A}, \mu)$ is closed. Because the operator $\mathscr{E}_{\mathfrak{B}}$ is contractive, it follows that $\mathscr{E}_{\mathfrak{B}}$ maps $L_2(X, \mathfrak{A}, \mu)$ into $L_2(X, \mathfrak{A}, \mu)$. We will denote the restriction of $\mathscr{E}_{\mathfrak{B}}$ to $L_2(X, \mathfrak{A}, \mu)$ by $\mathscr{E}_{\mathfrak{B}}|$. By definition, the range of the operator $\mathscr{E}_{\mathfrak{B}} : L_2(X, \mathfrak{A}, \mu) \to L_2(X, \mathfrak{A}, \mu)$ is a subspace of $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, but then by Proposition 7.29, the subspace $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ is fixed by $\mathscr{E}_{\mathfrak{B}}$, and so the range of $\mathscr{E}_{\mathfrak{B}}$ is actually equal to $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.

Remark 11.5. Some implications are used without much fanfare; namely, if the classes $\overline{f}^{\mu}$ and $\overline{g}^{\mu}$ are in $L_2(X, \mathfrak{A}, \mu)$, then the product $\overline{f}^{\mu}\overline{g}^{\mu} = \overline{fg}^{\mu}$ is in $L_1(X, \mathfrak{A}, \mu)$. This is a routine application of the Cauchy-Schwarz inequality, considered to be the $2 + 2$ of higher mathematics.

Proposition 11.6. *Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The conditional expectation operator $\mathscr{E}_{\mathfrak{B}} : L_2(X, \mathfrak{A}, \mu) \to L_2(X, \mathfrak{A}, \mu)$ is the unique orthogonal projection of the Hilbert space $L_2(X, \mathfrak{A}, \mu)$ onto the closed subspace $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.*

Proof. To show that $\mathscr{E}_{\mathfrak{B}}$ is the orthogonal projection of $L_2(X, \mathfrak{A}, \mu)$ onto $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, we need to show that to each $f \in L_2(X, \mathfrak{A}, \mu)$ there corresponds a unique element of $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, in this case $\mathscr{E}_{\mathfrak{B}} f$, such that $f - \mathscr{E}_{\mathfrak{B}} f \perp h$ for all $h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$.

We first show that if $f \in L_2(X, \mathfrak{A}, \mu)$, then $f - \mathscr{E}_{\mathfrak{B}} f \perp h$ for all $h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. Let $f \in L_2(X, \mathfrak{A}, \mu)$. We must show that

$$\langle f - \mathscr{E}_{\mathfrak{B}} f, h\rangle = 0 \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}}),$$

which is equivalent to showing that

$$\langle f, h\rangle = \langle \mathscr{E}_{\mathfrak{B}} f, h\rangle \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}}),$$

and in turn, this means showing that

$$\mathsf{E}\left(fh\right) = \mathsf{E}\left(h\,\mathscr{E}_{\mathfrak{B}}\,f\right) \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}}).$$

Let $h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. In order to apply Proposition 7.28, check first that $f \in L_1(X, \mathfrak{A}, \mu)$, and this is because $f \in L_2(X, \mathfrak{A}, \mu)$ and $L_2(X, \mathfrak{A}, \mu) \subseteq L_1(X, \mathfrak{A}, \mu)$ (since the measure space is finite). Next check that h is $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable since $h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. Finally check that $fh \in L_1(X, \mathfrak{A}, \mu)$ by Cauchy-Schwarz. Therefore, since $f \in L_1(X, \mathfrak{A}, \mu)$ and $h : X \to \mathbb{R}$ is $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable with $fh \in L_1(X, \mathfrak{A}, \mu)$, it follows by Proposition 7.28 that $\mathsf{E}\left(fh\right) = \mathsf{E}\left(h\,\mathscr{E}_{\mathfrak{B}}\,f\right)$, as required.

We now show the uniqueness part; namely we show that if $f \in L_2(X, \mathfrak{A}, \mu)$, then $\mathscr{E}_{\mathfrak{B}}\,f$ is the unique element of $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ such that $f - \mathscr{E}_{\mathfrak{B}}\,f \perp h$ for all $h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. Let $f \in L_2(X, \mathfrak{A}, \mu)$, and suppose that there is a $g \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ having the property that $f - g \perp h$ for all $h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$. We want to see that this implies g must equal $\mathscr{E}_{\mathfrak{B}}\,f$. To say that

$$f - g \perp h \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$$

means saying that

$$\langle f - g, h\rangle = 0 \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}}),$$

or equivalently that

$$\langle f, h\rangle = \langle g, h\rangle \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}}).$$

In terms of integrals, this says

$$\int_X fh\,d\mu = \int_X gh\,d\mu \quad \text{for all } h \in L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}}).$$

In particular, for every $B \in \mathfrak{B}$, the characteristic function χ_B is in $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, and so

$$\int_X f\chi_B \, d\mu = \int_X g\chi_B \, d\mu \quad \text{for all } B \in \mathfrak{B}.$$

Since $g\chi_B \in L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ for every $B \in \mathfrak{B}$, by Cauchy-Schwarz, it follows by Proposition 4.14 that

$$\int_X g\chi_B \, d\mu = \int_X g\chi_B \, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B}.$$

Consequently,

$$\int_X f\chi_B \, d\mu = \int_X g\chi_B \, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B},$$

which is the same as saying that

$$\int_B f \, d\mu = \int_B g \, d\mu_{|\mathfrak{B}} \quad \text{for all } B \in \mathfrak{B}.$$

This means that $g = \mathscr{E}_{\mathfrak{B}} f$; in fact, this is the defining feature of $\mathscr{E}_{\mathfrak{B}} f$, which completes the proof. ■

Example 11.7. Let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ be the smallest σ-subalgebra of $\mathfrak{A}$, namely $\{X, \varnothing\}$. Then the conditional expectation operator $\mathscr{E}_{\mathfrak{B}} : L_1(X, \mathfrak{A}, \mu) \to L_1(X, \mathfrak{A}, \mu)$ is equal to the expected value operator E times the class containing the constant function 1_X on $L_1(X, \mathfrak{A}, \mu)$, meaning $\mathscr{E}_{\mathfrak{B}}(\overline{g}^{\mu}) = \mathsf{E}\,\overline{g}^{\mu} \cdot \overline{1_X}^{\mu}$ for all $\overline{g}^{\mu} \in L_1(X, \mathfrak{A}, \mu)$. In this case, the closed subspace $L_2(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$ consists of all classes that are constant $\mu_{|\mathfrak{B}}^*$-almost everywhere on X. Recall Discussion 7.25.

12 Squared-Bias/Variance Decompositions

Throughout this section, let $(X, \mathfrak{A}, \mu)$ be a probability space, and let $\mathfrak{B}$ denote a σ-subalgebra of $\mathfrak{A}$.

Definition 12.1. Let $\mathscr{T}$ be a bounded operator on $L_1(X, \mathfrak{A}, \mu)$. If for every pair f and g in $L_1(X, \mathfrak{A}, \mu)$ such that $f \mathscr{T} g \in L_1(X, \mathfrak{A}, \mu)$, the equality

$$\mathscr{T}(f \mathscr{T} g) = (\mathscr{T} f)(\mathscr{T} g)$$

holds, then the operator $\mathscr{T}$ is called an **averaging operator**. Such an operator is said to satisfy the *averaging property*.

As will be shown in Proposition 12.4, every conditional expectation operator is an averaging operator.

Lemma 12.2. *If $\mathscr{T}$ is an averaging operator on $L_1(X, \mathfrak{A}, \mu)$, and if g is in the range of $\mathscr{T}$, then*

$$\mathscr{T}(fg) = (\mathscr{T} f)g = g \mathscr{T} f.$$

Proof. Let $\mathscr{T}$ be an averaging operator on $L_1(X, \mathfrak{A}, \mu)$, and let g be in the range of $\mathscr{T}$; that is, let $g = \mathscr{T} h$. Since $\mathscr{T}$ is an averaging operator, it follows by definition that $\mathscr{T}(f \mathscr{T} h) = \mathscr{T} f \mathscr{T} h$. Therefore, by substitution,

$$\mathscr{T}(fg) = \mathscr{T}(f \mathscr{T} h) = \mathscr{T} f \mathscr{T} h = (\mathscr{T} f)g = g \mathscr{T} f,$$

as required. ∎

Proposition 12.3. *Let $\mathscr{T}$ be an averaging operator on $L_1(X, \mathfrak{A}, \mu)$. If f and g are in $L_2(X, \mathfrak{A}, \mu)$ with $f \in \operatorname{Ker} \mathscr{T}$ and $g \in \operatorname{Ran} \mathscr{T}$, then*

$$\mathscr{T}\big((f+g)^2\big) = \mathscr{T}(f^2) + \mathscr{T}(g^2), \text{ and}$$
$$\mathscr{T}\big((f-g)^2\big) = \mathscr{T}(f^2) + \mathscr{T}(g^2).$$

Proof. Let f and g be in $L_2(X, \mathfrak{A}, \mu)$ with $f \in \operatorname{Ker} \mathscr{T}$ and $g \in \operatorname{Ran} \mathscr{T}$. The verification of the second displayed equality is short:

$$\begin{aligned}
\mathscr{T}\big((f-g)^2\big) &= \mathscr{T}(f^2 - 2fg + g^2) \\
&= \mathscr{T}(f^2) - 2\,\mathscr{T}(fg) + \mathscr{T}(g^2) \\
&= \mathscr{T}(f^2) - 2g\,\mathscr{T}(f) + \mathscr{T}(g^2) \quad g \in \operatorname{Ran} \mathscr{T}, \text{ Lemma 12.2} \\
&= \mathscr{T}(f^2) - 2g \cdot 0 + \mathscr{T}(g^2) \qquad f \in \operatorname{Ker} \mathscr{T} \\
&= \mathscr{T}(f^2) + \mathscr{T}(g^2),
\end{aligned}$$

and the verification of the first displayed equality amounts to replacing the minus sign with a plus sign, completing the proof. ■

Proposition 12.4. *If $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$, then $\mathscr{E}_{\mathfrak{B}}$ is an averaging operator. That is, if f and g are in $L_1(X, \mathfrak{A}, \mu)$, and if $\mathfrak{B}$ is a σ-subalgebra of $\mathfrak{A}$ with $f\,\mathscr{E}_{\mathfrak{B}}\,g \in L_1(X, \mathfrak{A}, \mu)$, then*

$$\mathscr{E}_{\mathfrak{B}}(f\,\mathscr{E}_{\mathfrak{B}}\,g) = (\mathscr{E}_{\mathfrak{B}}\,f)(\mathscr{E}_{\mathfrak{B}}\,g).$$

Proof. Since $\mathscr{E}_{\mathfrak{B}}\,g$ is $(\mathfrak{B}, \mathfrak{Bor}_{\mathbb{R}})$-measurable, and $f\,\mathscr{E}_{\mathfrak{B}}\,g \in L_1(X, \mathfrak{A}, \mu)$, it follows by Proposition 7.28 that $\mathscr{E}_{\mathfrak{B}}(f\,\mathscr{E}_{\mathfrak{B}}\,g) = (\mathscr{E}_{\mathfrak{B}}\,f)(\mathscr{E}_{\mathfrak{B}}\,g)$, as required. ■

Definition 12.5. Let f and g be in $L_2(X, \mathfrak{A}, \mu)$, and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$. The **conditional variance** of f given $\mathfrak{B}$, which we will denote by $\mathscr{V}ar_{\mathfrak{B}}(f)$, is defined by

$$\mathscr{V}ar_{\mathfrak{B}}(f) \coloneqq \mathscr{E}_{\mathfrak{B}}\big([f - \mathscr{E}_{\mathfrak{B}}\,f]^2\big).$$

Likewise, the (absolute, or unconditional) **variance** of f, which we will denote by $\mathsf{Var}\,(f)$, is defined by

$$\mathsf{Var}\,(f) \coloneqq \mathsf{E}\,\big((f - \mathsf{E}\,f)^2\big).$$

The **conditional covariance** of f and g given $\mathfrak{B}$, which we may denote by $\mathscr{C}ov_{\mathfrak{B}}(f,g)$, is defined by

$$\mathscr{C}ov_{\mathfrak{B}}(f,g) \coloneqq \mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}}\,f][g - \mathscr{E}_{\mathfrak{B}}\,g].$$

Likewise, the (absolute, or unconditional) **covariance** of f and g, which we may denote by $\mathsf{Cov}\,(f,g)$, is defined by

$$\mathsf{Cov}\,(f,g) \coloneqq \mathsf{E}\,\big((f - \mathsf{E}\,f)(g - \mathsf{E}\,g)\big).$$

Definition 12.6. Let f and g be in $L_1(X,\mathfrak{A},\mu)$. Inasmuch as f can be considered an estimator of the estimand g, the **conditional bias** in the estimator f given a σ-subalgebra $\mathfrak{B}$ is then $\mathscr{E}_{\mathfrak{B}}\,f - g$.

Remark 12.7. If a function f is a $(\mathfrak{B},\mathfrak{Bor}_{\mathbb{R}})$-measurable function, then so is its square f^2. Generally, the collection of $(\mathfrak{B},\mathfrak{Bor}_{\mathbb{R}})$-measurable functions is at least closed with respect to products. Consequently, if $f \in \operatorname{Ran}\mathscr{E}_{\mathfrak{B}}$, then $f^2 \in \operatorname{Ran}\mathscr{E}_{\mathfrak{B}}$.

Proposition 12.8. *Let $f \in L_2(X,\mathfrak{A},\mu)$ and let $\mathfrak{B}$ be a σ-subalgebra of $\mathfrak{A}$ with $g \in L_2(X,\mathfrak{B},\mu_{|\mathfrak{B}})$. Then*

$$\begin{aligned}\mathscr{E}_{\mathfrak{B}}[f-g]^2 &= \underbrace{\mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}}\,f]^2}_{\textit{cond'l variance}} + \underbrace{[\mathscr{E}_{\mathfrak{B}}\,f - g]^2}_{\textit{squared cond'l bias}} \\ &= \mathscr{V}ar_{\mathfrak{B}}(f) + [\mathscr{E}_{\mathfrak{B}}\,f - g]^2\end{aligned}$$

Proof. In order to apply Proposition 12.3, we acknowledge that $\mathscr{E}_{\mathfrak{B}}$ is an averaging operator (Proposition 12.4) and that $f - \mathscr{E}_{\mathfrak{B}}\,f \in \operatorname{Ker}\mathscr{E}_{\mathfrak{B}}$ and that $\mathscr{E}_{\mathfrak{B}}\,f - g \in \operatorname{Ran}\mathscr{E}_{\mathfrak{B}}$. By Proposition 12.3, it follows that

$$\mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}}\,f + \mathscr{E}_{\mathfrak{B}}\,f - g]^2 = \mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}}\,f]^2 + \mathscr{E}_{\mathfrak{B}}[\mathscr{E}_{\mathfrak{B}}\,f - g]^2.$$

Since $\mathscr{E}_{\mathfrak{B}} f - g \in \operatorname{Ran} \mathscr{E}_{\mathfrak{B}}$, it follows that $[\mathscr{E}_{\mathfrak{B}} f - g]^2 \in \operatorname{Ran} \mathscr{E}_{\mathfrak{B}}$. And a projection fixes its range, so

$$\mathscr{E}_{\mathfrak{B}}[\mathscr{E}_{\mathfrak{B}} f - g]^2 = [\mathscr{E}_{\mathfrak{B}} f - g]^2.$$

Consequently, combining these equalities,

$$\begin{aligned} \mathscr{E}_{\mathfrak{B}}[f - g]^2 &= \mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}} f + \mathscr{E}_{\mathfrak{B}} f - g]^2 \\ &= \mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}} f]^2 + \mathscr{E}_{\mathfrak{B}}[\mathscr{E}_{\mathfrak{B}} f - g]^2 \\ &= \mathscr{E}_{\mathfrak{B}}[f - \mathscr{E}_{\mathfrak{B}} f]^2 + [\mathscr{E}_{\mathfrak{B}} f - g]^2, \end{aligned}$$

as required. ■

13 Acknowledgements

- Most of the measure and integration theory came from *Principle of Real Analysis* [2], Aliprantis and Burkinshaw.
- The conditional distribution material started with Loeve, *Probability Theory* [5].
- The Bayesian statistics here started as an elaboration of sections 1.2.1 and 1.2.2 in *Elements of Bayesian Statistics* [3], Florens, Mouchart, and Rolin.
- The Bayesian statistics examples largely came from *Statistical Rethinking* [6], Richard McElreath.
- The abstract algebra can be found in any elementary algebra text. Hungerford, say.

14 Afterword

Bibliography

[1] Le Chang and Doris Y. Tsao. The code for facial identity in the primate brain. *Cell*, 169:1013–1028.e14, 2017.

[2] Charalambos D. Aliprantis and Owen Burkinshaw. *Principle of Real Analysis*. Academic Press, Second edition, 1990.

[3] Jean-Pierre Florens, Michel Mouchart, and Jean-Marie Rolin. *Elements of Bayesian Statistics*. Dekker, 1990.

[4] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea, 1956.

[5] Michel Loeve. *Probability Theory*. Van Nostrand, Second edition, 1960.

[6] Richard McElreath. *Statistical Rethinking*. CRC Press, Second edition, 2020.

Nomenclature

$\mathfrak{Bor}_X$ Borel sets of the topological space X, page 13

$\mathscr{E}_{\mathfrak{B}}\,\overline{f}$ conditional expectation in $L_1(X, \mathfrak{A}, \mu)$, page 68

$E_{\mathfrak{B}}\,\overline{f}$ conditional expectation in $L_1(X, \mathfrak{B}, \mu_{|\mathfrak{B}})$, page 69

$\mathfrak{C} \perp\!\!\!\perp \mathfrak{D} \mid \mathfrak{B}$, the σ-algebras $\mathfrak{C}$ and $\mathfrak{D}$ are conditionally independent given $\mathfrak{B}$, page 113

μ-$\mathfrak{Meas}$, the σ-algebra of μ-measurable subsets, where μ is an outer measure, page 7

μ^* outer measure on the power set, or its restriction to the μ^*-measurable subsets, page 7

$\sigma(\mathfrak{S}, \mathfrak{T})$ product σ-algebra generated by $\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$, page 44

$\mathfrak{Semi}(\mathfrak{S}, \mathfrak{T})$ semiring of Cartesian products $A \times B$ with $A \in \mathfrak{S}$ and $B \in \mathfrak{T}$, page 42

Index