36-752 Advanced Probability Overview Spring 2018

6. Conditional Probability and Expectation

Instructor: Alessandro Rinaldo

Associated reading: Chapter 5 of Ash and Doléans-Dade; Sec 5.1 of Durrett.

Overview

In this set of lecture notes we shift focus to dependent random variables. We introduce measure-theoretic definitions of conditional probability and conditional expectations.

1 Conditional Expectation

The measure-theoretic definition of conditional expectation is a bit unintuitive, but we will show how it matches what we already know from earlier study.

Definition 1 (Conditional Expectation). *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\mathcal{C} \subseteq \mathcal{F}$ be a sub-σ-field. Let X be a random variable that is $\mathcal{F}/\mathcal{B}^1$ measurable and $\mathrm{E}|X| < \infty$. We use the symbol $\mathrm{E}(X|\mathcal{C})$ to stand for any function $h : \Omega \to \mathbb{R}$ that is $\mathcal{C}/\mathcal{B}^1$ measurable and that satisfies*

$$\int_C h dP = \int_C X dP, \text{ for all } C \in \mathcal{C}. \tag{1}$$

We call such a function h, a version *of the* conditional expectation of X given $\mathcal{C}$.

Equation (1) can also be written $\mathrm{E}(I_C h) = \mathrm{E}(I_C X)$ for all $C \in \mathcal{C}$. Any two versions of $\mathrm{E}(X|\mathcal{C})$ must be equal a.s. according to Theorem 21 in Lecture Notes Set 3 (part 3). Also, any $\mathcal{C}/\mathcal{B}^1$-measurable function that equals a version of $\mathrm{E}(X|\mathcal{C})$ a.s. is another version.

Example 2. *If X is itself $\mathcal{C}/\mathcal{B}^1$ measurable, then X is a version of $\mathrm{E}(X|\mathcal{C})$.*

Example 3. *If $X = a$ a.s., then $\mathrm{E}(X|\mathcal{C}) = a$ a.s.*

Let Y be a random quantity and let $\mathcal{C} = \sigma(Y)$. We will use the notation $\mathrm{E}(X|Y)$ to stand for $\mathrm{E}(X|\mathcal{C})$. According to Theorem 39 (given in the Appendix), $\mathrm{E}(X|Y)$ is some function $g(Y)$ because it is $\sigma(Y)/\mathcal{B}^1$-measurable. We will also use the notation $\mathrm{E}(X|Y = y)$ to stand for $g(y)$.

Example 4 (Joint Densities). *Let (X, Y) be a pair of random variables with a joint density $f_{X,Y}$ with respect to Lebesgue measure. Let $\mathcal{C} = \sigma(Y)$. The usual marginal and conditional densities are*

$$\begin{aligned} f_Y(y) &= \int f_{X,Y}(x,y)dx, \\ f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)}. \end{aligned}$$

The traditional calculation of the conditional mean of X given $Y = y$ is

$$g(y) = \int x f_{X|Y}(x|y)dx.$$

That is, $\mathrm{E}(X|Y) = g(Y)$ is the traditional definition of conditional mean of X given Y. We also use the symbol $\mathrm{E}(X|Y = y)$ to stand for $g(y)$. We can prove that $h = g(Y)$ is a version of the conditional mean according to Definition 1. Since $g(Y)$ is a function of Y, we know that it is $\mathcal{C}/\mathcal{B}^1$ measurable. We need to show that Equation (1) holds. Let $C \in \mathcal{C}$ so that there exists $B \in \mathcal{B}^1$ so that $C = Y^{-1}(B)$. Then $I_C(\omega) = I_B(Y(\omega))$ for all ω. Then

$$\begin{aligned} \int_C h dP &= \int I_C h dP \\ &= \int I_B(Y) g(Y) dP \\ &= \int I_B g d\mu_Y \\ &= \int I_B(y) g(y) f_Y(y) dy \\ &= \int I_B(y) \int x f_{X|Y}(x|y) dx f_Y(y) dy \\ &= \int\int I_B(y) x f_{X,Y}(x,y) dx dy \\ &= \mathrm{E}(I_B(Y)X) = \mathrm{E}(I_C X). \end{aligned}$$

Example 4 can be extended easily to handle two more general cases. First, we could find $\mathrm{E}(r(X)|Y)$ by virtually the same calculation. Second, the use of conditional densities extends to the case in which the joint distribution of (X, Y) has a density with respect to an arbitrary product measure.

All of the familiar results about conditional expectation are special cases of the general definition. Here is an unfamiliar example.

Example 5. *Let X_1, X_2 be independent with $U(0, \theta)$ distribution for some known θ. Let $Y = \max\{X_1, X_2\}$ and $X = X_1$. We want to find the conditional mean of X given Y.*

Intuitively, with probability 1/2, $X = Y$, and with probability 1/2, X is the min of X_1 and X_2 and ought to be uniformly distributed between 0 and Y. The mean of this hybrid distribution is $Y/2 + Y/4 = 3Y/4$. Let's verify this.

First, we see that $h = 3Y/4$ is measurable with respect to $\mathcal{C} = \sigma(Y)$. Next, let $C \in \mathcal{C}$. We need to show that $\mathrm{E}(XI_C) = \mathrm{E}([3Y/4]I_C)$. Theorem 21 of Lecture Notes Set 3 (part 4) says that we only need to check this for sets of the form $C = Y^{-1}([0, d])$ with $0 < d < \theta$. Rewrite these expectations as integrals with respect to the joint distribution of (X_1, X_2). We need to show that

$$\int_0^d \int_0^d \frac{x_1}{\theta^2} dx_1 dx_2 = \int_0^d \frac{3y}{4} \frac{2y}{\theta^2} dy, \tag{2}$$

for all $0 < d < \theta$. It is easy to see that both sides of Equation (2) equal $d^3/[2\theta^2]$. Furthermore,

$$h' = \begin{cases} 3Y/4 & \text{if } Y \text{ is irrational,} \\ 0 & \text{otherwise.} \end{cases}$$

is another version of $\mathrm{E}(X|Y)$.

Here is a simple property that extends from expectations to conditional expectations. It will be used to prove the existence of conditional expectations.

Lemma 6 (Monotonicity). *If $X_1 \leq X_2$ a.s., then $\mathrm{E}(X_1|\mathcal{C}) \leq \mathrm{E}(X_2|\mathcal{C})$ a.s.*

Proof: Suppose that both $\mathrm{E}(X_1|\mathcal{C})$ and $\mathrm{E}(X_2|\mathcal{C})$ exist. Let

$$\begin{aligned} C_0 &= \{\infty > \mathrm{E}(X_1|\mathcal{C}) > \mathrm{E}(X_2|\mathcal{C})\}, \\ C_1 &= \{\infty = \mathrm{E}(X_1|\mathcal{C}) > \mathrm{E}(X_2|\mathcal{C})\}. \end{aligned}$$

Then, for $i = 0, 1$,

$$0 \leq \int_{C_i} [\mathrm{E}(X_1|\mathcal{C}) - \mathrm{E}(X_2|\mathcal{C})] dP = \int_{C_i} (X_1 - X_2) dP \leq 0.$$

It follows that all terms in this string are 0 and $P(C_i) = 0$ for $i = 0, 1$. Since $C_0 \cup C_1 = \{\mathrm{E}(X_1|\mathcal{C}) > \mathrm{E}(X_2|\mathcal{C})\}$, the result is proven. ■

2 Existence of Conditional Expectation

We could prove that versions of conditional expectations exist by the Radon-Nikodym theorem. However, the "modern" way to prove the existence of conditional expectations is through the theory of Hilbert spaces.

Definition 7. *An* inner product space *is a vector space $\mathcal{V}$ with an inner product $\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, a function that satisfies*

- *symmetry:* $\langle u, v \rangle = \langle v, u \rangle$,
- *bilinearity (part 1):* $\langle u_1 + u_2, v \rangle = \langle u_1, v \rangle + \langle u_2, v \rangle$,
- *bilinearity (part 2): for real* λ, $\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$,
- *positivity:* $\langle u, u \rangle > 0$ *for all* $u \neq 0$, *and* $\langle u, u \rangle = 0$ *if and only if* $u = 0$.

An inner product provides a norm, namely $\|u\| = \sqrt{\langle u, u \rangle}$ and a metric $d(u, v) = \|u - v\|$. These facts follow from some simple properties of inner products.

Proposition 8. *Let $\mathcal{V}$ be a vector space with an inner product $\langle \cdot, \cdot \rangle$. Then*

1. **Parallelogram law:** *for all $u, v \in \mathcal{V}$, $\|u\|^2 + \|v\|^2 = \frac{1}{2}(\|u + v\|^2 + \|u - v\|^2)$.*
2. **Cauchy-Schwarz inequality:** *for all $u, v \in \mathcal{V}$, $|\langle u, v \rangle| \leq \|u\| \|v\|$, with equality if and only if u and v are collinear.*
3. **Triangle inequality:** *for all $u, v \in \mathcal{V}$, $\|u + v\| \leq \|u\| + \|v\|$.*

Definition 9. *A complete (see Definition 7 in Lecture Notes Set 6) inner product space is a* Hilbert space.

Example 10. *Let $\mathcal{V} = L^2(\Omega, \mathcal{F}, \mu)$. Define $\langle f, g \rangle = \int f g d\mu$. This is an inner product that produces the norm $\| \cdot \|_2$. Lemma 9 of Lecture Notes Set 6 showed that L^p is complete.*

We prove existence of conditional expectations using orthogonal projection in Hilbert spaces. The following theorem is a basic result in Hilbert space theory, and is proved in the Appendix.

Theorem 11 (Existence and uniqueness of orthogonal projections). *Let $\mathcal{V}$ be a Hilbert space and let $\mathcal{V}_0$ be a closed subspace. For each $v \in \mathcal{V}$, there is a unique $v_0 \in \mathcal{V}_0$ (called the* orthogonal projection of v into $\mathcal{V}_0$*) such that $v - v_0$ is orthogonal to every vector in $\mathcal{V}_0$ and $\|v - v_0\| = \inf_{w \in \mathcal{V}_0} \|v - w\|$.*

Now, we can prove the existence of conditional expectations.

Theorem 12 (Existence of conditional expectation). *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let Y be a random variable. Let $\mathcal{C}$ be a sub-σ-field of $\mathcal{F}$. If $\mathrm{E}(Y)$ exists, then there exists a version of $\mathrm{E}(Y|\mathcal{C})$.*

Proof: It is easy to see that $L^2(\Omega, \mathcal{C}, P)$ is a closed linear subspace. If $Y \in L^2(\Omega, \mathcal{F}, P)$, let Y_0 be the projection of Y into $L^2(\Omega, \mathcal{C}, P)$. According to Theorem 11, $\mathrm{E}([Y - Y_0]X) = 0$ for all $X \in L^2(\Omega, \mathcal{C}, P)$, in particular for $X = I_C$ for arbitrary $C \in \mathcal{C}$.

If $Y > 0$ but not in L^2, define $Y_n = \min\{Y, n\}$. Then $Y_n \in L^2$. Let $Y_{0,n}$ be a version of $\mathrm{E}(Y_n|\mathcal{C})$, and assume that $Y_{0,n} \leq Y_{0,n+1}$ for all n, which is allowed by Lemma 6. Let $Y_0 = \lim_{n\to\infty} Y_{n,0}$, which exists (by monotonicity of the conditional expectation) and is $\mathcal{C}$-measurable. Thus, for each $C \in \mathcal{C}$,

$$\begin{aligned} \mathrm{E}(I_C Y) &= \lim_{n\to\infty} \mathrm{E}(I_C Y_n), \\ &= \lim_{n\to\infty} \mathrm{E}(I_C Y_{0,n}), \\ &= \mathrm{E}(I_C Y_0), \end{aligned}$$

by the conditional and unconditional version of the monotone convergence theorem. It follows that $\mathrm{E}(I_C Y) = \mathrm{E}(I_C Y_0)$ for all $C \in \mathcal{C}$ and Y_0 is a version of $\mathrm{E}(Y|\mathcal{C})$.

If Y takes both positive and negative values, write $Y = Y^+ - Y^-$. If one of the means $\mathrm{E}(Y^+)$ or $\mathrm{E}(Y^-)$ is finite then the probability is 0 that both $\mathrm{E}(Y^+|\mathcal{C}) = \infty$ and $\mathrm{E}(Y^-|\mathcal{C}) = \infty$. Then $\mathrm{E}(Y^+|\mathcal{C}) - \mathrm{E}(Y^-|\mathcal{C})$ is a version of $\mathrm{E}(Y|\mathcal{C})$. ∎

The following result summarizes what we have learned about the existence and uniqueness of conditional expectation.

Corollary 13. *If $Y \in L^1(\Omega, \mathcal{F}, P)$ and $\mathcal{C}$ is a sub-σ-field of $\mathcal{F}$. Let $Z \in L^2(\Omega, \mathcal{C}, P)$, then the following are equivalent.*

1. *$Z = \mathrm{E}(Y|\mathcal{C})$.*
2. *$\mathrm{E}(XZ) = \mathrm{E}(XY)$ for all $X \in L^2(\Omega, \mathcal{C}, P)$.*
3. *Z is the orthogonal projection of Y into $L^2(\Omega, \mathcal{C}, P)$.*

3 Additional Properties of Conditional Expectation

The following fact is immediate by letting $\mathcal{C} = \mathcal{F}$.

Proposition 14. $\mathrm{E}(\mathrm{E}(X|\mathcal{C})) = \mathrm{E}(X)$.

Here is a generalization of Proposition 14, which is sometimes called the *tower property* of conditional expectations, or *law of total probability*.

Proposition 15 (William's Tower Property). *If $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \mathcal{F}$ are sub-σ-field's and $\mathrm{E}(X)$ exists, then $\mathrm{E}(X|\mathcal{C}_1)$ is a version of $\mathrm{E}(\mathrm{E}(X|\mathcal{C}_2)|\mathcal{C}_1)$.*

Proof: By definition $\mathrm{E}(X|\mathcal{C}_1)$ is $\mathcal{C}_1/\mathcal{B}^1$-measurable. We need to show that, for every $C \in \mathcal{C}_1$,

$$\int_C \mathrm{E}(X|\mathcal{C}_1)dP = \int_C \mathrm{E}(X|\mathcal{C}_2)dP.$$

The left side is $\mathrm{E}(XI_C)$ by definition of conditional mean. Similarly, because $C \in \mathcal{C}_2$ also, the right side is $\mathrm{E}(XI_C)$ as well. ■

Example 16. *Let (X, Y, Z) be a triple of random variables. Then $\mathrm{E}(X|Y)$ is a version of $\mathrm{E}(\mathrm{E}(X|(Y, Z))|Y)$.*

The following corollary to Proposition 15 is sometimes useful.

Corollary 17. *Assume that $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \mathcal{F}$ are sub-σ-field's and $\mathrm{E}(X)$ exists. If a version of $\mathrm{E}(X|\mathcal{C}_2)$ is $\mathcal{C}_1/\mathcal{B}^1$-measurable, then $\mathrm{E}(X|\mathcal{C}_1)$ is a version of $\mathrm{E}(X|\mathcal{C}_2)$ and $\mathrm{E}(X|\mathcal{C}_2)$ is a version of $\mathrm{E}(X|\mathcal{C}_1)$.*

Example 18. *Suppose that X and Y have a joint conditional density given Θ that factors,*

$$f_{X,Y|\Theta}(x, y|\theta) = f_{X|\Theta}(x|\theta)f_{Y|\Theta}(y|\theta).$$

Then, the conditional density of X given (Y, Θ) is

$$f_{X|Y,\Theta}(x|y, \theta) = \frac{f_{X,Y|\Theta}(x, y|\theta)}{f_{Y|\Theta}(y|\theta)} = f_{X|\Theta}(x|\theta).$$

With $\mathcal{C}_1 = \sigma(\Theta)$ and $\mathcal{C}_2 = \sigma(Y, \Theta)$, we see that $\mathrm{E}(r(X)|\mathcal{C}_1)$ will be a version of $\mathrm{E}(r(X)|\mathcal{C}_2)$ for every function $r(X)$ with defined mean.

The next lemma shows that conditional expectation is linear.

Lemma 19 (Linearity). *If $\mathrm{E}(X)$, $\mathrm{E}(Y)$, and $\mathrm{E}(X + Y)$ all exist, then $\mathrm{E}(X|\mathcal{C}) + \mathrm{E}(Y|\mathcal{C})$ is a version of $\mathrm{E}(X + Y|\mathcal{C})$.*

Proof: Clearly $\mathrm{E}(X|\mathcal{C}) + \mathrm{E}(Y|\mathcal{C})$ is $\mathcal{C}/\mathcal{B}^1$-measurable. We need to show that for all $C \in \mathcal{C}$,

$$\int_C \mathrm{E}(X|\mathcal{C}) + \mathrm{E}(Y|\mathcal{C})dP = \int_C (X + Y)dP. \tag{3}$$

The left side of Equation (3) is $\int_C XdP + \int_C YdP = \int_C (X + Y)dP$ because $\mathrm{E}(I_C X)$, $\mathrm{E}(I_C Y)$ and $\mathrm{E}(I_C[X + Y])$ all exist. ■

The following theorem is used extensively in later results.

Theorem 20. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{C}$ be a sub-σ-field of $\mathcal{F}$. Suppose that $\mathrm{E}(Y)$ and $\mathrm{E}(XY)$ exist and that X is $\mathcal{C}/\mathcal{B}^1$-measurable. Then $\mathrm{E}(XY|\mathcal{C}) = X\mathrm{E}(Y|\mathcal{C})$.*

Proof: Clearly, $X\mathrm{E}(Y|\mathcal{C})$ is $\mathcal{C}/\mathcal{B}^1$-measurable. We will use the standard machinery on X. If $X = I_B$ for a set $B \in \mathcal{C}$, then

$$\mathrm{E}(I_C XY) = \mathrm{E}(I_{C\cap B}Y) = \mathrm{E}(I_{C\cap B}\mathrm{E}(Y|\mathcal{C})) = \mathrm{E}(I_C X\mathrm{E}(Y|\mathcal{C})), \tag{4}$$

for all $C \in \mathcal{C}$. Hence, $X\mathrm{E}(Y|\mathcal{C}) = \mathrm{E}(XY|\mathcal{C})$. By linearity of expectation, the extreme ends of Equation (4) are equal for every nonnegative simple function, X. Next, suppose that X is nonnegative and let $\{X_n\}$ be a sequence of nonnegative simple functions converging to X from below. Then

$$\begin{aligned} \mathrm{E}(I_C X_n Y^+) &= \mathrm{E}(I_C X_n \mathrm{E}(Y^+|\mathcal{C})), \\ \mathrm{E}(I_C X_n Y^-) &= \mathrm{E}(I_C X_n \mathrm{E}(Y^-|\mathcal{C})), \end{aligned}$$

for each n and each $C \in \mathcal{C}$. Apply the monotone convergence theorem to all four sequences above to get

$$\begin{aligned} \mathrm{E}(I_C XY^+) &= \mathrm{E}(I_C X\mathrm{E}(Y^+|\mathcal{C})), \\ \mathrm{E}(I_C XY^-) &= \mathrm{E}(I_C X\mathrm{E}(Y^-|\mathcal{C})), \end{aligned}$$

for all $C \in \mathcal{C}$. It now follows easily from Lemma 19 that $X\mathrm{E}(Y|\mathcal{C}) = \mathrm{E}(XY|\mathcal{C})$. Finally, if X is general, use what we just proved to see that $X^+\mathrm{E}(Y|\mathcal{C}) = \mathrm{E}(X^+Y|\mathcal{C})$ and $X^-\mathrm{E}(Y|\mathcal{C}) = \mathrm{E}(X^-Y|\mathcal{C})$. Apply Lemma 19 one last time. ■

In all of the proofs so far, we have proven that the defining equation for conditional expectation holds for all $C \in \mathcal{C}$. Sometimes, this is too difficult and the following result can simplify a proof.

Proposition 21. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{C}$ be a sub-σ-field of $\mathcal{F}$. Let $\mathcal{D}$ be a π-system that generates $\mathcal{C}$. Let Y be a random variable whose mean exists. Let Z be a $\mathcal{C}/\mathcal{B}^1$-measurable random variable such that $\mathrm{E}(I_C Z) = \mathrm{E}(I_C Y)$ for all $C \in \mathcal{D}$. Then Z is a version of $\mathrm{E}(Y|\mathcal{C})$.*

One proof of this result relies on signed measures, and is very similar to the proof of uniqueness of measure.

4 Conditional Distribution

Now we introduce the measure-theoretic version of conditional probability and distribution.

4.1 Conditional probability

For $A \in \mathcal{F}$, define $\Pr(A|\mathcal{C}) = \mathrm{E}(I_A|\mathcal{C})$. That is, treat I_A as a random variable X and define the conditional probability of A to be the conditional mean of X. We would like to show that conditional probabilities behave like probabilities. The first thing we can show is that they are additive. That is a consequence of the following result.

It follows easily from Lemma 19 that $\Pr(A|\mathcal{C}) + \Pr(B|\mathcal{C}) = \Pr(A \cup B|\mathcal{C})$ a.s. if A and B are disjoint. The following additional properties are straightforward, and we will not do them all in class. They are similar to Lemma 19.

Example 22 (Probability at most 1). *We shall show that* $\Pr(A|\mathcal{C}) \leq 1$ *a.s. Let* $B = \{\omega : \Pr(A|\mathcal{C}) > 1\}$. *Then* $B \in \mathcal{C}$, *and*

$$P(B) \leq \int_B \Pr(A|\mathcal{C}) dP = \int_B I_A dP = P(A \cap B) \leq P(B),$$

where the first inequality is strict if $P(B) > 0$. *Clearly, neither of the inequalities can be strict, hence* $P(B) = 0$.

Example 23 (Countable Additivity). *Let* $\{A_n\}_{n=1}^{\infty}$ *be disjoint elements of* $\mathcal{F}$. *Let* $W = \sum_{n=1}^{\infty} \Pr(A_n|\mathcal{C})$. *We shall show that* W *is a version of* $\Pr(\bigcup_{n=1}^{\infty} A_n|\mathcal{C})$. *Let* $C \in \mathcal{C}$.

$$\begin{aligned}
\mathrm{E}\left[I_C I_{\cup_{n=1}^{\infty} A_n}\right] &= P\left(C \cap \left[\bigcup_{n=1}^{\infty} A_n\right]\right) \\
&= \sum_{n=1}^{\infty} P\left(C \cap A_n\right) \\
&= \sum_{n=1}^{\infty} \int_C \Pr(A_n|\mathcal{C}) dP \\
&= \int_C \sum_{n=1}^{\infty} \Pr(A_n|\mathcal{C}) dP \\
&= \int_C W dP,
\end{aligned}$$

where the sum and integral are interchangeable by the monotone convergence theorem.

We could also prove that $\Pr(A|\mathcal{C}) \geq 0$ a.s. and $\Pr(\Omega|\mathcal{C}) = 1$, a.s. But there are generally uncountably many different $A \in \mathcal{F}$ and uncountably many different sequences of disjoint events. Although countable additivity holds a.s. separately for each sequence of disjoint events, how can we be sure that it holds simultaneously for all sequences a.s.?

Definition 24 (Regular Conditional Probabilities). *Let* $\mathcal{A} \subseteq \mathcal{F}$ *be a sub-*σ*-field. We say that a collection of versions* $\{\Pr(A|\mathcal{C}) : A \in \mathcal{A}\}$ *are* regular conditional probabilities *if, for each* ω, $\Pr(\cdot|\mathcal{C})(\omega)$ *is a probability measure on* $(\Omega, \mathcal{A})$.

Rarely do regular conditional probabilities exist on $(\Omega, \mathcal{F})$, but there are lots of common sub-σ-field's $\mathcal{A}$ such that regular conditional probabilities exist on $(\Omega, \mathcal{A})$. Oddly enough, the existence of regular conditional probabilities doesn't seem to depend on $\mathcal{C}$.

Example 25 (Joint Densities). *Use the same setup as in Example 4. For each y such that $f_Y(y) = 0$, define $f_{X|Y}(x|y) = \phi(x)$, the standard normal density. For each y such that $f_Y(y) > 0$, define $f_{X|Y}$ as in Example 4. Next, for each $B \in \mathcal{B}^1$, let $A = X^{-1}(B)$, and define*

$$h(y) = \int_B f_{X|Y}(x|y)dx,$$

for all y. Finally, define $\Pr(A|\mathcal{C})(\omega) = h(Y(\omega))$. The calculation done in Example 4 shows that this is a version of the conditional mean of I_A given $\mathcal{C}$. But it is easy to see that for each ω, $\Pr(\cdot|\mathcal{C})(\omega)$ is a probability measure on $(\Omega, \sigma(X))$.

The results we have on existence of regular conditional probabilities are for the cases in which $\mathcal{A}$ is the σ-field generated by a random variable or something a lot like a random variable. Note that this is a condition on $\mathcal{A}$ *not* on $\mathcal{C}$. The conditioning σ-field can be anything at all. What matters is the σ-field on which the conditional probability is to be defined. There are examples in which no regular conditional probabilities exist. These examples depend upon the existence of a nonmeasurable set, which we did not prove. We will not cover such examples here.

4.2 Conditional distribution

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let $X : \Omega \to \mathcal{X}$ be a random quantity. If $\mathcal{A} = \sigma(X)$, conditional probabilities on $\mathcal{A}$ form a conditional distribution for X.

Definition 26 (Conditional distribution). *For each $B \in \mathcal{B}$, define $\mu_{X|\mathcal{C}}(B)(\omega) = \Pr(X^{-1}(B)|\mathcal{C})(\omega)$. A collection of versions $\{\mu_{X|\mathcal{C}}(B)(\cdot) : B \in \mathcal{B}\}$ is called a* conditional distribution of X given $\mathcal{C}$. *If, in addition, for each ω, $\mu_{X|\mathcal{C}}(\cdot)(\omega)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$, then the collection is a* regular conditional distribution (rcd).

Example 25 is already an example of an rcd.

Here is a bit of notation that we will use when we deal with conditional distributions given random quantities.

Definition 27 (Conditional distribution given a random quantity). *Let X and Y be random quantities (defined on the same probability space) taking values in $\mathcal{X}$ (with σ-field $\mathcal{B}$) and $\mathcal{Y}$ (with σ-field $\mathcal{D}$) respectively. Assume that $\mathcal{D}$ contains all singletons. We will use $\mu_{X|Y}$ to denote the conditional distribution of X given Y with the following meaning. For every $y \in \mathcal{Y}$ and $\omega \in Y^{-1}(\{y\})$ and $B \in \mathcal{B}$, we define $\mu_{X|Y}(B|y) = \mu_{X|\mathcal{C}}(B)(\omega)$, where $\mathcal{C} = \sigma(Y)$.*

Remember that a $\mathcal{C}/\mathcal{B}^1$-measurable function like $\mu_{X|\mathcal{C}}(B)$ must itself be a function of Y. That is, for all ω, $\mu_{X|\mathcal{C}}(B)(\omega) = h(Y(\omega))$, where $h : \mathcal{Y} \to \mathbb{R}$ is $\mathcal{D}/\mathcal{B}^1$-measurable. What we have done is define $\mu_{X|Y}(B|y) = h(y)$. We also use the notation $\Pr(X \in B|Y = y)$ to stand for this same value.

Here is a result that says that, in the presence of an rcd, conditional expectations can be computed the naïve way.

Proposition 28 (Expectation under RCD). *Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let $X : \Omega \to \mathcal{X}$ be a random quantity. Let $g : \mathcal{X} \to \mathbb{R}$ be a measurable function such that the mean of $g(X)$ exists. Suppose that $\mu_{X|\mathcal{C}}$ is an rcd for X given $\mathcal{C}$. Then $\int g d\mu_{X|\mathcal{C}}$ is a version of $\mathrm{E}(g(X)|\mathcal{C})$.*

Just use the standard machinery to prove this result.

Very often, a conditional distribution of X given Y is proposed on the space $\mathcal{X}$ in which X takes its values, and is given as a function of Y. To check whether such a proposed distribution is the conditional distribution of X given Y, the following result can help.

Lemma 29. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(\mathcal{X}, \mathcal{B})$ and $(\mathcal{Y}, \mathcal{D})$ be measurable spaces. Let $\mathcal{B}_1$ and $\mathcal{D}_1$ be π-systems that generate $\mathcal{B}$ and $\mathcal{D}$ respectively. Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ be random quantities. Let μ_Y stand for the distribution of Y and let $\mu_{X,Y}$ stand for the joint distribution of (X, Y). For each $B \in \mathcal{B}$, let $h_B : \mathcal{Y} \to \mathbb{R}$ be a measurable function such that for all $D \in \mathcal{D}_1$ and $B \in \mathcal{B}_1$, $\int_D h_B d\mu_Y = \mu_{X,Y}(B \times D)$. Then $\{h_B(Y) : B \in \mathcal{B}\}$ is a version of the conditional distribution of X given Y.*

Proof: Let $\mathcal{C} = \sigma(Y)$. Each $h_B(Y)$ is a $\mathcal{C}/\mathcal{B}^1$-measurable function from Ω to $\mathbb{R}$. We need to show that for all $C \in \mathcal{C}$ and $B \in \mathcal{B}$,

$$\int_C h_B(Y) dP = P(C \cap X^{-1}(B)). \tag{5}$$

First, notice that $\mathcal{F}_1 = \{C \cap X^{-1}(B) : C \in \mathcal{C}, B \in \mathcal{B}\}$ is a sub-σ-field of $\mathcal{F}$ and that both sides of Equation (5) define σ-finite measures on $(\Omega, \mathcal{F}_1)$. We will prove that these two measures agree on the π-system $\{Y^{-1}(D) \cap X^{-1}(B) : D \in \mathcal{D}_1, B \in \mathcal{B}_1\}$, which generates $\mathcal{F}_1$. Then apply the uniqueness of measure. For $D \in \mathcal{D}_1$ and $B \in \mathcal{B}_1$, let $C = Y^{-1}(D)$. It follows that

$$\int_C h_B(Y) dP = \int_D h_B d\mu_Y = \mu_{X,Y}(B \times D) = P(C \cap X^{-1}(B)).$$

■

Theorem 30 (Existence of RCD for r.v.'s). *Let $(\Omega, \mathcal{F}, P)$ be a probability space with $\mathcal{C}$ a sub-σ-field. Let X be a random variable. Then there is a rcd of X given $\mathcal{C}$.*

Proof: Define

$$\begin{aligned} C_1 &= \left\{\omega : \mu_{X|\mathcal{C}}((-\infty, q])(\omega) = \inf_{\text{rational } r>q} \mu_{X|\mathcal{C}}((-\infty, r])(\omega), \text{ for all rational } q\right\}, \\ C_2 &= \left\{\omega : \lim_{x\to-\infty,\ x \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, x])(\omega) = 0\right\}, \\ C_3 &= \left\{\omega : \lim_{x\to\infty,\ x \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, x])(\omega) = 1\right\}. \end{aligned}$$

Let $C_0 = C_1 \cap C_2 \cap C_3$. In another course document, we give details on why $P(C_0) = 1$. For $\omega \in C_0$ and irrational x, define

$$\mu_{X|\mathcal{C}}((-\infty, x])(\omega) = \inf_{\text{rational } q>x} \mu_{X|C}((-\infty, q])(\omega). \tag{6}$$

Rational x already satisfy Equation (6) for $\omega \in C_0$ since $C_0 \subseteq C_1$. If $\omega \in C_0^C$, define $\mu_{X|\mathcal{C}}((-\infty, x])(\omega) = F(x)$, where F is your favorite cdf. For each ω, we have defined a cdf on $\mathbb{R}$, which extends to a probability measure on $(\mathbb{R}, \mathcal{B}^1)$. This collection of probabilities forms an rcd by construction. ■

Further discussions are given in the Appendix.

The following result says that RCD exists under bimeasurable mappings.

Lemma 31. *Let $X : \Omega \to \mathcal{X}$ and let $\phi : \mathcal{X} \to R \in \mathcal{B}^1$ be on-to-one, onto, measurable, with measurable inverse. Let $Y = \phi(X)$. Let $\mu_{Y|\mathcal{C}}$ be an rcd for Y given $\mathcal{C}$. Define $\mu_{X|\mathcal{C}}(B)(\omega) = \mu_{Y|\mathcal{C}}(\phi(B))(\omega)$. Then $\mu_{X|\mathcal{C}}$ defines an rcd for X given $\mathcal{C}$.*

The proof of Lemma 31, along with some examples, are given in the Appendix.

There is a laundry list of properties of conditional expectation that mimic similar properties of expectation. Proposition 28 can be used to prove some of these. Using Proposition 28 requires the existence of an rcd. Theorem 30 shows that an rcd exists for every random variable.

Proposition 32 (Integral properties under RCD). *Let $\mathcal{C}$ be a sub-σ-field of $\mathcal{F}$.*

1. (Monotone convergence theorem) *If $0 \le X_n \le X$ a.s. for all n and $X_n \to X$ a.s., then $\mathrm{E}(X_n|\mathcal{C}) \to \mathrm{E}(X|\mathcal{C})$ a.s.*
2. (Dominated convergence theorem) *If $X_n \to X$ a.s. and $|X_n| \le Y$ a.s., where $Y \in L^1$, then $\mathrm{E}(X_n|\mathcal{C}) \to \mathrm{E}(X|\mathcal{C})$ a.s.*
3. (Jensen's inequality) *Let $\mathrm{E}(X)$ be finite. If ϕ is a convex function and $\phi(X) \in L^1$, then $\mathrm{E}[\phi(X)|\mathcal{C}] \ge \phi(\mathrm{E}[X|\mathcal{C}])$ a.s.*

4. *Assume that* $\mathrm{E}(X)$ *exists and* $\sigma(X)$ *is independent of* $\mathcal{C}$. *Then* $\mathrm{E}(X)$ *is a version of* $\mathrm{E}(X|\mathcal{C})$.

Proof: We will prove only the first part. Let $X_0 = X$ and let $\boldsymbol{X} = (X_0, X_1, X_2, \ldots)$. Let $\mu_{\boldsymbol{X}|\mathcal{C}}(\cdot)$ be a regular conditional distribution for $\boldsymbol{X}$ given $\mathcal{C}$. That is, for each $B \in \mathcal{B}^\infty$ (the product σ-field for $\mathbb{R}^\infty$) $\mu_{\boldsymbol{X}|\mathcal{C}}(B)(\cdot)$ (a function from Ω to $\mathbb{R}$) is a version of $\Pr(\boldsymbol{X} \in B|\mathcal{C})(\cdot)$, and for each $\omega \in \Omega$, $\mu_{\boldsymbol{X}|\mathcal{C}}(\cdot)(\omega)$ is a probability measure on $\mathcal{B}^\infty$. Let $L = \{\boldsymbol{x} \in \mathbb{R}^\infty : \lim_{n\to\infty} x_n = x_0\}$. Let $C = \{\omega : \mu_{\boldsymbol{X}|\mathcal{C}}(L)(\omega) = 1\}$. We know that

$$\begin{aligned} 1 &= P(\boldsymbol{X} \in L) \\ &= \int \mu_{\boldsymbol{X}|\mathcal{C}}(L)dP \\ &= \int_C \mu_{\boldsymbol{X}|\mathcal{C}}(L)dP + \int_{C^C} \mu_{\boldsymbol{X}|\mathcal{C}}(L)dP \\ &= P(C) + \int_{C^C} \mu_{\boldsymbol{X}|\mathcal{C}}(L)dP, \end{aligned}$$

where the first equality follows from $\lim_{n\to\infty} X_n = X_0$ a.s., the second follows from the law of total probability Proposition 14, and the last two are obvious. If $P(C) < 1$, the last integral above is strictly less than $1 - P(C)$ contradicting the first equality. Hence $P(C) = 1$.[1] Define $f_n(\boldsymbol{x}) = x_n$ for $n = 0, 1, \ldots$. For each $\omega \in C$, $\lim_{n\to\infty} f_n = f_0$, a.s. $[\mu_{\boldsymbol{X}|\mathcal{C}}(\cdot)(\omega)]$. The monotone convergence theorem says that, for each $\omega \in C$,

$$\lim_{n\to\infty} \int f_n \mu_{\boldsymbol{X}|\mathcal{C}}(d\boldsymbol{x})(\omega) = \int f_0 \mu_{\boldsymbol{X}|\mathcal{C}}(d\boldsymbol{x})(\omega).$$

According to Proposition 28,

$$\begin{aligned} \int f_n(\boldsymbol{x})\mu_{\boldsymbol{X}|\mathcal{C}}(d\boldsymbol{x})(\omega) &= \mathrm{E}(f_n(\boldsymbol{X})|\mathcal{C}) = \mathrm{E}(X_n|\mathcal{C}), \\ \int f_0(\boldsymbol{x})\mu_{\boldsymbol{X}|\mathcal{C}}(d\boldsymbol{x})(\omega) &= \mathrm{E}(f_0(\boldsymbol{X})|\mathcal{C}) = \mathrm{E}(X_0|\mathcal{C}). \end{aligned}$$

It follows that $\lim_{n\to\infty} \mathrm{E}(X_n|\mathcal{C}) = \mathrm{E}(X|\mathcal{C})$ a.s. ∎

5 Bayes' Theorem

Let $(\mathcal{X}, \mathcal{B})$ be a Borel space and let $X : \Omega \to \mathcal{X}$ be a random quantity. Also, let $\Theta : \Omega \to \mathcal{T}$, where $(\mathcal{T}, \tau)$ is a measurable space. We can safely assume that X has an rcd given Θ. Let ν be a measure on $(\mathcal{X}, \mathcal{B})$ such that $\mu_{X|\Theta}(\cdot|\theta)$ has a density $f_{X|\Theta}(x|\theta)$ with respect to ν for all θ. Assume that $f_{X|\Theta}$ is jointly measurable as a function of (x, θ).

[1] This illustrates a common technique in probability proofs. We integrate a nonnegative function $f \leq 1$ with respect to a probability P and find that $\int f dP = 1$. It follows that $P(f = 1) = 1$.

Theorem 33 (Bayes' theorem). *Assume the above structure. Then*

$$f_X(x) = \int_{\mathcal{T}} f_{X|\Theta}(x|\theta) d\mu_\Theta(\theta) \text{ is a density for } \mu_X \text{ with respect to } \nu. \tag{7}$$

Also, $\mu_{\Theta|X}(\cdot|x) \ll \mu_\Theta$ *a.s.* $[\mu_X]$ *and the following is a density for* $\mu_{\Theta|X}$ *with respect to* μ_Θ*:*

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)}{f_X(x)}.$$

Surprisingly, Bayes' theorem does not require that $(\mathcal{T}, \tau)$ be a Borel space. Nevertheless, we get an rcd for Θ given X. The proof of Theorem 33 is given in another course document.

6 Independence and Conditioning

Recall that $\{X_\alpha\}_{\alpha\in\aleph}$ are mutually independent if, for all finite k and all distinct $i_1, \ldots, i_k \in \aleph$, the joint distributions satisfy

$$\mu_{X_{i_1},\ldots,X_{i_k}} = \mu_{X_{i_1}} \times \cdots \times \mu_{X_{i_k}}.$$

Theorem 34. *Random quantities* $\{X_\alpha\}_{\alpha\in\aleph}$ *are mutually independent if and only if for all integers* k_1, k_2 *(such that* $k_1 + k_2$ *is no more than the cardinality of* $\aleph$*) and distinct* $i_1, \ldots, i_{k_1}, j_1, \ldots, j_{k_2} \in \aleph$*,* $\mu_{X_{j_1},\ldots,X_{j_{k_2}}}$ *is a version of* $\mu_{X_{j_1},\ldots,X_{j_{k_2}}|X_{i_1},\ldots,X_{i_{k_1}}}$*.*

Proof: For the "if" direction, we shall use induction. Start with $k = 2$, $i_1 \neq j_1 = i_2$, and $k_1 = k_2 = 1$. We assume that $\mu_{X_{j_1}}$ is a version of $\mu_{X_{j_1}|X_{i_1}}$, so

$$\mu_{X_{i_1},X_{i_2}}(B_1 \times B_2) = \int_{B_1} \mu_{X_{i_2}}(B_2) d\mu_{X_{i_1}} = \mu_{X_{i_1}}(B_1)\mu_{X_{i_2}}(B_2).$$

So, X_{i_1} and X_{i_2} are independent for all distinct i_1 and i_2. Now, assume that the "if" implication is true for all $k \leq k_0$. Let $i_1, \ldots, i_{k_0+1}$ be distinct elements of $\aleph$. Let $k_1 = k_0$ and $k_2 = 1$. Let $j_1 = i_{k_0+1}$. Then $\mu_{X_{i_{k_0+1}}}$ is a version of $\mu_{X_{i_{k_0+1}}|X_{i_1},\ldots,X_{i_{k_0}}}$. Using the same argument as above, we we see that $X_{i_1}, \ldots, X_{i_{k_0+1}}$ are independent.

For the "only if" direction, assume that the random variables are independent. Let $i_1, \ldots, i_{k_1}, j_1, \ldots, j_{k_2}$ be distinct. Let B_1 be in the product σ-field of the spaces where $X_{i_1}, \ldots, X_{i_{k_1}}$ take their values and let B_2 be in the product σ-field of the spaces where $X_{j_1}, \ldots, X_{j_{k_2}}$ take their values. We have assumed that

$$\begin{aligned} \mu_{X_{i_1},\ldots,X_{j_{k_2}}}(B_1 \times B_2) &= \mu_{X_{i_1},\ldots,X_{i_{k_1}}}(B_1)\mu_{X_{j_1},\ldots,X_{j_{k_2}}}(B_2) \\ &= \int_{B_1} \mu_{X_{j_1},\ldots,X_{j_{k_2}}}(B_2) d\mu_{X_{i_1},\ldots,X_{i_{k_1}}}. \end{aligned}$$

This equality for all B_1 and B_2 is sufficient (by Lemma 29) to say that $\mu_{X_{j_1},\ldots,X_{j_{k_2}}}$ is a version of $\mu_{X_{j_1},\ldots,X_{j_{k_2}}|X_{i_1},\ldots,X_{i_{k_1}}}$. ∎

Theorem 35. *Two σ-field's $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent if and only if, for every $\mathcal{C}_1$-measurable random variable X such that $\mathrm{E}(X)$ is defined, $\mathrm{E}(X)$ is a version of $\mathrm{E}(X|\mathcal{C}_2)$.*

Proof: We know that $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent if and only if, for all $A_i \in \mathcal{C}_i$ $(i = 1, 2)$ $P(A_1 \cap A_2) = P(A_1)P(A_2)$. Notice that $P(A_1) = \mathrm{E}(I_{A_1})$ is a version of $\mathrm{E}(I_{A_1}|\mathcal{C}_2) = \Pr(A_1|\mathcal{C}_2)$ if and only if,

$$P(A_1 \cap A_2) = \int_{A_2} P(A_1) dP, \quad \text{for all } A_2 \in \mathcal{C}_2.$$

But the right side equals $P(A_1)P(A_2)$. Hence, we have proven that $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent if and only if $\mathrm{E}(I_{A_1})$ is a version of $\mathrm{E}(I_{A_1}|\mathcal{C}_2)$ for all $A_1 \in \mathcal{C}_1$. The extension to all $\mathcal{C}_1$-measurable random variables is an application of the standard machinery (part 4 of Proposition 32, which you are proving for homework). ■

The following definition is useful in statistics.

Definition 36. *We say that $X_1, X_2, \ldots$ are* conditionally independent *given $\mathcal{C}$ if either of the following conditions holds:*

- *For all k_1, k_2 and distinct $i_1, \ldots, i_{k_1}, j_1, \ldots, j_{k_2}$, $\mu_{X_{j_1},\ldots,X_{j_{k_2}}|\mathcal{C}}$ is a version of $\mu_{X_{j_1},\ldots,X_{j_{k_2}}|\sigma(\mathcal{C},X_{i_1},\ldots,X_{i_{k_1}})}$*
- *For every k and distinct $i_1, \ldots, i_k$, $\mu_{X_{i_1}|\mathcal{C}} \times \cdots \times \mu_{X_{i_k}|\mathcal{C}}$ is a version of $\mu_{X_{i_1},\ldots,X_{i_k}|\mathcal{C}}$*

Proposition 37. *The two conditions in the above definition are equivalent.*

Example 38. *A sequence $\{X_n\}_{n=1}^{\infty}$ is called a* Markov chain *if, for all $n > 1$, $(X_1, \ldots, X_{n-1})$ is conditionally independent of $\{X_k\}_{k=n+1}^{\infty}$ given X_n.*

Appendix

A Conditional expectation given a random variable

Theorem 39. *Let $(\Omega_i, \mathcal{F}_i)$ for $i = 1, 2, 3$ be measurable spaces. Let $f : \Omega_1 \to \Omega_2$ be a measurable onto function. Suppose that $\mathcal{F}_3$ contains all singletons. Let $\mathcal{A}_1 = \sigma(f)$. Let $g : \Omega_1 \to \Omega_3$ be $\mathcal{F}_1/\mathcal{F}_3$-measurable. Then g is $\mathcal{A}_1/\mathcal{F}_3$-measurable if and only if there exists a $\mathcal{F}_2/\mathcal{F}_3$-measurable $h : \Omega_2 \to \Omega_3$ such that $g = h(f)$.*

Proof: For the "if" part, assume that there is a measurable $h : \Omega_2 \to \Omega_3$ such that $g(\omega) = h(f(\omega))$ for all $\omega \in \Omega_1$. Let $B \in \mathcal{F}_3$. We need to show that $g^{-1}(B) \in \mathcal{A}_1$. Since h is measurable, $h^{-1}(B) \in \mathcal{F}_2$, so $h^{-1}(B) = A$ for some $A \in \mathcal{F}_2$. Since $g^{-1}(B) = f^{-1}(h^{-1}(B))$, it follows that $g^{-1}(B) = f^{-1}(A) \in \mathcal{A}_1$.

For the "only if" part, assume that g is $\mathcal{A}_1$ measurable. For each $t \in \Omega_3$, let $C_t = g^{-1}(\{t\})$. Since g is measurable with respect to $\mathcal{A}_1 = f^{-1}(\mathcal{F}_2)$, every element of $g^{-1}(\mathcal{F}_3)$ is in $f^{-1}(\mathcal{F}_2)$. So let $A_t \in \mathcal{F}_2$ be such that $C_t = f^{-1}(A_t)$. Define $h(\omega) = t$ for all $\omega \in A_t$. (Note that if $t_1 \neq t_2$, then $A_{t_1} \cap A_{t_2} = \varnothing$, so h is well defined.) To see that $g(\omega) = h(f(\omega))$, let $g(\omega) = t$, so that $\omega \in C_t = f^{-1}(A_t)$. This means that $f(\omega) \in A_t$, which in turn implies $h(f(\omega)) = t = g(\omega)$.

To see that h is measurable, let $A \in \mathcal{F}_3$. We must show that $h^{-1}(A) \in \mathcal{F}_2$. Since g is $\mathcal{A}_1$ measurable, $g^{-1}(A) \in \mathcal{A}_1$, so there is some $B \in \mathcal{F}_2$ such that $g^{-1}(A) = f^{-1}(B)$. We will show that $h^{-1}(A) = B \in \mathcal{F}_2$ to complete the proof. If $\omega \in h^{-1}(A)$, let $t = h(\omega) \in A$ and $\omega = f(x)$ (because f is onto). Hence, $x \in C_t \subseteq g^{-1}(A) = f^{-1}(B)$, so $f(x) \in B$. Hence, $\omega \in B$. This implies that $h^{-1}(A) \subseteq B$. Lastly, if $\omega \in B$, $\omega = f(x)$ for some $x \in f^{-1}(B) = g^{-1}(A)$ and $h(\omega) = h(f(x)) = g(x) \in A$. So, $h(\omega) \in A$ and $\omega \in h^{-1}(A)$. This implies $B \subseteq h^{-1}(A)$. ∎

The condition that f be onto can be relaxed at the expense of changing the domain of h to be the image of f, i.e. $h : f(\Omega_1) \to \Omega_3$, with a different σ-field. The proof is slightly more complicated due to having to keep track of the image of f, which might not be a measureable set in $\mathcal{F}_2$.

The following is an example to show why the condition that $\mathcal{F}_3$ contains all singletons is included in Theorem 39.

Example 40. *Let $\Omega_i = \mathbb{R}$ for all i and let $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{B}^1$, while $\mathcal{F}_3 = \{\mathbb{R}, \varnothing\}$. Then every function $g : \Omega_1 \to \Omega_3$ is $\sigma(f)/\mathcal{F}_3$-measurable, no matter what $f : \Omega_1 \to \Omega_2$ is. For example, let $f(x) = x^2$ and $g(x) = x$ for all x. Then $g^{-1}(\mathcal{F}_3) \subseteq \sigma(f)$ but g is not a function of f.*

The reason that we need the condition about singletons is the following. Suppose that there are two points $t_1, t_2 \in \Omega_3$ such that $t_1 \in A$ implies $t_2 \in A$ and vice versa for every $A \in \mathcal{F}_3$. Then there can be a set $A \in \mathcal{F}_3$ that contains both t_1 and t_2, and g can take both of the values t_1 and t_2, but f is constant on $g^{-1}(A)$ and all the measurability conditions still hold. In this case, g is not a function of f.

B Projection in Hilbert spaces

Theorem 11 is well known in finite dimensional spaces. The following lemma aids in the general proof.

Lemma 41. *Let x_1, x_2, x be elements of an inner product space. Then*

$$\|x_1 - x_2\|^2 = 2\|x_1 - x\|^2 + 2\|x_2 - x\|^2 - 4\|(x_1 + x_2)/2 - x\|^2. \tag{8}$$

Proof: Use the relation between inner products and norms to compute

$$\begin{array}{rcl}
\|x_1 - x_2\|^2 & = & \langle x_1, x_1\rangle + \langle x_2, x_2\rangle - 2\langle x_1, x_2\rangle, \\
2\|x_1 - x\|^2 & = & 2\langle x_1, x_1\rangle + 2\langle x, x\rangle - 4\langle x_1, x\rangle, \\
2\|x_2 - x\|^2 & = & 2\langle x_2, x_2\rangle + 2\langle x, x\rangle - 4\langle x_2, x\rangle, \\
-4\|(x_1 + x_2)/2 - x\|^2 & = & -\langle x_1, x_1\rangle - \langle x_2, x_2\rangle - 2\langle x_1, x_2\rangle - 4\langle x, x\rangle + 4\langle x_1, x\rangle + 4\langle x_2, x\rangle.
\end{array}$$

Add the last three rows, and the sum is the first row. ∎

Proof: [Proof of Theorem 11] Fix $v \in \mathcal{V}$. Define $g(w) = \|w - v\|^2$ and let $c_0 = \inf_{w \in \mathcal{V}_0} g(w)$. Let $\{v_n\}_{n=1}^{\infty}$ be elements of $\mathcal{V}_0$ such that $\lim_{n\to\infty} g(v_n) = c_0$. We will prove that $\{v_n\}_{n=1}^{\infty}$ is a Cauchy sequence. If not, there is a subsequence, call it $\{y_n\}_{n=1}^{\infty}$, and $\epsilon > 0$ such that $\|y_n - y_{n+1}\| > \epsilon$ for all n. For each n, use Lemma 41 with $x_1 = y_n$, $x_2 = y_{n+1}$, $x = v$ to conclude that, for all n,

$$\begin{array}{rcl}
\epsilon^2 & < & \|y_n - y_{n+1}\|^2 \\
& = & 2\|y_n - v\|^2 + 2\|y_{n+1} - v\|^2 - 4\|(y_n + y_{n+1})/2 - v\|^2. \qquad (9)
\end{array}$$

Notice that $\limsup_n \|y_n - v\|^2 = \limsup_n \|y_{n+1} - v\|^2 = c_0^2$ and $\liminf_n \|(y_n + y_{n+1})/2 - v\|^2 \geq c_0^2$. It follows that the $\limsup_n$ of the far right of Appendix B is at most 0, a contradiction. Because $\mathcal{V}_0$ is complete, it follows that $\{v_n\}_{n=1}^{\infty}$ has a limit v_0. Because g is continuous, $\|v_0 - v\| = c_0$.

Next, let $w \in \mathcal{V}_0$ be nonzero and define $c = \langle v_0 - v, w\rangle$. Notice that

$$\|aw + v_0 - v\|^2 = \|v_0 - v\|^2 + 2ac + a^2\|w\|^2 \geq \|v_0 - v\|^2,$$

by the definition of v_0. It follows that $h(a) = 2ac + a^2\|w\|^2 \geq 0$ for all a. But the function h has a unique minimum of $-c^2/\|w\|^2$ at $a = -c/\|w\|^2$, hence $c = 0$. So $v_0 v$ is orthogonal to every vector in $\mathcal{V}_0$.

Finally, show that v_0 is unique. Suppose that there is v_1 such that $\|v - v_1\| = \|v - v_0\|$. Apply Lemma 41 with $x_1 = v_0$, $x_2 = v_1$, and $x = v$. The left side of Equation (8) is nonnegative while the right side is nonpositive, so they are both 0 and $v_1 = v_0$. ∎

C More on existence of RCD

We give further detailed discussion of the existence of RCD.

C.1 More on the proof of Theorem 30

For each rational number q, let $\mu_{X|\mathcal{C}}((-\infty, q])$ be a version of $\Pr(X \leq q|\mathcal{C})$. Define

$$\begin{aligned} C_1 &= \left\{\omega : \mu_{X|\mathcal{C}}((-\infty, q])(\omega) = \inf_{\text{rational } r>q} \mu_{X|\mathcal{C}}((-\infty, r])(\omega), \text{ for all rational } q\right\}, \\ C_2 &= \left\{\omega : \lim_{x\to-\infty,\ x \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, x])(\omega) = 0\right\}, \\ C_3 &= \left\{\omega : \lim_{x\to\infty,\ x \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, x])(\omega) = 1\right\}. \end{aligned}$$

(Notice that C_2 and C_3 are defined slightly differently than in the original class notes.)
Define

$$M_{q,r} = \{\omega : \mu_{X|\mathcal{C}}((-\infty, q](\omega) < \mu_{X|\mathcal{C}}((-\infty, r])(\omega))\}, \quad M = \bigcup_{q>r} M_{q,r}.$$

If $P(M_{q,r}) > 0$, for some $q > r$ then

$$\begin{aligned} \Pr(M_{q,r} \cap \{X \leq q\}) &= \int_{M_{q,r}} \mu_{X|\mathcal{C}}((-\infty, q])dP < \int_{M_{q,r}} \mu_{X|\mathcal{C}}((-\infty, q])dP \\ &= \Pr(M_{q,r} \cap \{X \leq r\}), \end{aligned}$$

which is a contradiction. Hence, $P(M) = 0$. Next, define

$$N_q = \{\omega \in M^C : \lim_{r\downarrow q,\ r \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, r](\omega) \neq \mu_{X|\mathcal{C}}((-\infty, q])(\omega), \quad N = \bigcup_{\text{All q}} N_q.$$

If $P(N_q) > 0$ for some q, then

$$\begin{aligned} \Pr(N_q \cap \{X \leq q\}) &= \int_{N_q} \mu_{X|\mathcal{C}}((-\infty, q])dP < \int_{N_q} \lim_{r\downarrow q,\ r \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, r])dP \\ &= \lim_{r\downarrow q,\ r \text{ rational}} \int \mu_{X|\mathcal{C}}((-\infty, r])dP = \lim_{r\downarrow q,\ r \text{ rational}} \Pr(N_q \cap \{X \leq r\}), \end{aligned}$$

which is a contradiction. We can use Example 23 once again to prove that $P(N) = 0$. Notice that $C_1 = N^C$, so $P(C_1) = 1$.
Next, notice that

$$\begin{aligned} 0 = P\left(C_1 \cap C_2^C \cap \bigcap_{\text{rational } x} \{X \leq x\}\right) &= \lim_{x\to-\infty,\ x \text{ rational}} \int_{C_1\cap C_2^C} \mu_{X|\mathcal{C}}((-\infty, x])dP \\ &= \int_{C_1\cap C_2^C} \lim_{x\to-\infty,\ x \text{ rational}} \mu_{X|\mathcal{C}}((-\infty, x])dP. \end{aligned}$$

If $P(C_1 \cap C_2) < 1$, then the last integral above is strictly positive, a contradiction. The interchange of limit and integral is justified by the fact that, for $\omega \in C_1$, $\mu_{X|\mathcal{C}}((-\infty, x])$ is nondecreasing in x. A similar contradiction arises if $P(C_1 \cap C_3) < 1$.

C.2 Bimeasurable functions and Borel spaces

Definition 42 (Bimeasurable functions). *A one-to-one onto measurable function ϕ between two measurable spaces is called* bimeasurable *if its inverse is also measurable. Suppose that there exists a bimeasurable $\phi : \mathcal{X} \to R$, where R is a measurable subset of $\mathbb{R}$. In this case, we say that $(\mathcal{X}, \mathcal{B})$ is a* Borel space.

Lemma 43. *Let $X : \Omega \to \mathcal{X}$ and let $\phi : \mathcal{X} \to R \subseteq \mathbb{R}$ be bimeasurable, where $R \in \mathcal{B}^1$. Let $Y = \phi(X)$. Let $\mu_{Y|\mathcal{C}}$ be an rcd for Y given $\mathcal{C}$. Define $\mu_{X|\mathcal{C}}(B)(\omega) = \mu_{Y|\mathcal{C}}(\phi(B))(\omega)$. Then $\mu_{X|\mathcal{C}}$ defines an rcd for X given $\mathcal{C}$.*

Proof: Recall that ϕ^{-1} is measurable, hence, for each ω, $\mu_{X|\mathcal{C}}(\cdot)(\omega) = \mu_{Y|\mathcal{C}}(\phi(\cdot))(\omega)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$. For each $B \in \mathcal{B}$, $\mu_{X|\mathcal{C}}(B)(\cdot)$ is a measurable function of ω. What remains is to verify that for all $C \in \mathcal{C}$ and $B \in \mathcal{B}$, $P(C \cap X^{-1}(B)) = \int_C \mu_{X|\mathcal{C}}(B) dP$. But the left side of this is

$$P(C \cap Y^{-1}(\phi(B))) = \int_C \mu_{Y|\mathcal{C}}(\phi(B)) dP = \int_C \mu_{X|\mathcal{C}}(B) dP.$$

■

Besides $(\mathbb{R}, \mathcal{B}^1)$ and measurable subspaces, what else are Borel spaces?

- Finite and countable products of Borel spaces are Borel spaces.
- Complete separable metric spaces (Polish spaces) are Borel spaces.
- The collection of all continuous functions on the closed unit interval with the L^∞ norm and Borel σ-field is a Borel space.

These results are proven in Section B.3.2 of Schervish (1995) *Theory of Statistics*. Here is one example.

Example 44. (Bimeasurable Function from $(0,1)^\infty$ to a Subset of $(0,1)$) *For each $x \in (0,1)$, define*

$$\begin{aligned} y_0(x) &= x, \\ z_j(x) &= \begin{cases} 1 & \text{if } 2y_{j-1}(x) \geq 1, \\ 0 & \text{if not} \end{cases} \quad \text{for } j = 1, 2, \ldots, \\ y_j(x) &= 2y_{j-1}(x) - z_j(x), \text{ for } j = 1, 2, \ldots, \end{aligned}$$

This makes $z_j(x)$ the jth bit in the binary expansion of x that has infinitely many 0's. Also, each z_j is a measurable function. Construct the following array that contains each positive

integer once and only once:

$$\begin{array}{cccccc} 1 & 3 & 6 & 10 & 15 & \cdots \\ 2 & 5 & 9 & 14 & 20 & \cdots \\ 4 & 8 & 13 & 19 & 26 & \cdots \\ 7 & 12 & 18 & 25 & 33 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

Let $\ell(i,j)$ stand for the jth number from the top of the ith column. Now, define

$$\phi(x_1, x_2, \ldots) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{z_j(x_i)}{2^{\ell(i,j)}}.$$

Intuitively, ϕ takes each x_i and places its binary expansion down column i of a doubly infinite array and then combines all the bits in the order of the array above into a single number. It is easy to see that ϕ is a limit of measurable functions, so it is measurable. Its inverse is $\phi^{-1} = (g_1, g_2, \ldots)$ where

$$g_i(x) = \sum_{j=1}^{\infty} \frac{z_{\ell(i,j)}(x)}{2^j}.$$

This definition makes $z_j(g_i(x)) = z_{\ell(i,j)}(x)$, confirming that $\phi(g_1(x), g_2(x), \ldots) = x$. Also, each g_i is measurable, so ϕ^{-1} is measurable. The range R of the function ϕ is all elements of $(0,1)$ except $\bigcup_{i=1}^{\infty} B_i$, where B_i is defined as follows. For each finite subset I of $\{\ell(i,j)\}_{j=1}^{\infty}$, let $C_{i,I}$ be the set of all x that have 1's in the bits of their binary expansions corresponding to all coordinates not in I. Then $B_i = \bigcup_I C_{i,I}$. Since there are only countably many finite subsets I of each $\{\ell(i,j)\}_{j=1}^{\infty}$, B_i is a countable union of sets. Each $C_{i,I}$ is measurable, so R is measurable.

One can extend Example 44 to $\mathbb{R}^{\infty}$ by first mapping $\mathbb{R}^{\infty}$ to $(0,1)^{\infty}$ using a strictly increasing cdf on each coordinate. Also, a slightly simpler argument can map $(0,1)^k$ into $(0,1)$, so that $\mathbb{R}^k$ is a Borel space. Indeed, the argument in Example 44 proves that products of Borel spaces are Borel spaces. (Just map each Borel space into $\mathbb{R}$ first and then into $(0,1)$ and then apply the example to the product.)