MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2018
Lecture 11

CONTINUOUS RANDOM VARIABLES - II

Contents

1 REVIEW OF JOINT DISTRIBUTIONS

Recall that two random variables X and Y are said to be jointly continuous if there exists a nonnegative measurable function $f_{X,Y}$ such that

$$\mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)\, dv\, du.$$

Once we have in our hands a general definition of integrals, this can be used to establish that for every Borel subset of $\mathbb{R}^2$, we have

$$\mathbb{P}((X,Y) \in B) = \int_B f_{X,Y}(u,v)\, du\, dv.$$

Furthermore, X is itself a continuous random variable, with density f_X given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy.$$

Finally, recall that $\mathbb{E}[g(X)] = \int g(x) f_X(x)\, dx$. Similar to the discrete case, the expectation of $g(X) = X^m$ and $g(X) = (X - \mathbb{E}[X])^m$ is called the mth moment and the mth central moment, respectively, of X. In particular, $\mathrm{var}(X) \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2]$ is the variance of X.

We note that all of the definitions and formulas have obvious extensions to the case of more than two random variables.

2 MARKOV KERNELS

Random variables X and Y endowed with a product measures $\mathbb{P}_X \times \mathbb{P}_Y$ are necessarily independent $X \perp\!\!\!\perp Y$. How do we construct $\mathbb{P}_{X,Y}$ for dependent variables? One method is to define X and Y on the same probability space and compute $\mathbb{P}_{X,Y}$ using the Definition given in Lecture 10. Another method involves the following concept:

Definition 1. *$K : \Omega_1 \times \mathcal{F}_2 \to [0, 1]$ is called a transition probability kernel (or a Markov kernel) acting from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$ if:*

1. *$K(\omega_1, \cdot)$ is a probability measure on $(\Omega_2, \mathcal{F}_2)$ for each $\omega_1 \in \Omega_1$*
2. *$\omega_1 \mapsto K(\omega_1, B)$ is an $\mathcal{F}_1$-measurable function for each $B \in \mathcal{F}_2$.*

In some disciplines, it is common to abuse notation and say "Let $K : (\Omega_1, \mathcal{F}_1) \to (\Omega_2, \mathcal{F}_2)$ be a Markov kernel" or even "Let $K : \Omega_1 \to \Omega_2$ be a Markov kernel", even though K is not a map between spaces.

Example: When Ω_1 and Ω_2 are finite, any Markov kernel K acting from $(\Omega_1, 2^{\Omega_1})$ to $(\Omega_2, 2^{\Omega_2})$ is simply an $|\Omega_1| \times |\Omega_2|$ matrix of non-negative values with row-sums all equal to 1. Such matrices are called stochastic (or right-stochastic, or row-stochastic).

Example: The following transition probability kernel acts between $(\mathbb{R}, \mathcal{B})$ and $(\mathbb{R}, \mathcal{B})$. It is called the *additive Gaussian noise channel*:

$$K(x, dy) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}}\, dy \quad x, y \in \mathbb{R}.$$

This kernel "blurs" every point into a Gaussian cloud of width σ.

Theorem 1. *For any probability measure $\mathbb{P}_1$ and transition probability kernel K there exists a unique probability measure π (denoted $\mathbb{P}_1 \times K$) on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ such that*

$$\pi(A \times B) = \int_A K(\omega_1, B)\mathbb{P}_1(d\omega_1).$$

Furthermore, whenever $f \geq 0$ or f is π-integrable we have

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d\pi = \int_{\Omega_1} \mathbb{P}_1(d\omega_1) \int_{\Omega_2} f(\omega_1, \omega_2) K(\omega_1, d\omega_2)\,. \quad (1)$$

Proof. Repeat the steps in the proofs of Theorems 2 and 3 in Lecture 9 with trivial modifications. □

The measure π on $\Omega_1 \times \Omega_2$ corresponds to the following stochastic experiment:

- Draw ω_1 in accordance with distribution $\mathbb{P}_1(\cdot)$.
- Then draw ω_2 in accordance with distribution $K(\omega_1, \cdot)$.
- Output pair (ω_1, ω_2).

Caution: Many different kernels can lead to the same product measure, i.e.

$$\mathbb{P}_1 \times K = \mathbb{P}_1 \times K' \quad \not\Rightarrow K = K'.$$

Indeed if $\mathbb{P}_1(A) = 0$, then $K(x, \cdot)$ can be defined arbitrarily for all $x \in A$ without affecting the product measure.

2.1 Measure-kernel-function

Markov kernels can act on functions and on measures and these actions are associative.

Proposition 1. *Let K be a Markov kernel from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$. Then*

1. *The kernel K pulls back non-negative functions f on Ω_2 to non-negative functions on Ω_1:*

$$(Kf)(\omega_1) \triangleq \int_{\Omega_2} f(\omega_2) K(\omega_1, d\omega_2),$$

and this map $\omega_1 \mapsto (Kf)(\omega_1)$ is $\mathcal{F}_1$ measurable.

2. *The kernel K pushes forward probability measures from Ω_1 to Ω_2. Namely for each μ on $(\Omega_1, \mathcal{F}_1)$ there exists a unique probability measure $\nu = \mu K$ on $(\Omega_2, \mathcal{F}_2)$ satisfying*

$$\nu(B) = \int_{\Omega_1} K(\omega_1, B) d\mu. \tag{2}$$

3. *These actions are compatible: for any μ on Ω_1 and $f \geq 0$ on Ω_2*

$$\int_{\Omega_1} (Kf)(\omega_1)\, d\mu = \int_{\Omega_2} f(\omega_2)\, d\nu\,. \tag{3}$$

Proof. We only sketch the details. For 1 notice that measurability of Kf for simple functions follows from the definition of Markov kernel. This extends to general functions by taking limits. For 2 notice that by the MCT assignment (2) indeed defines a σ-additive probability measure. Finally, 3 is obvious for simple functions and otherwise take limits. □

It is common to denote the integral $\int f d\mu$ as μf or $\mu(f)$, i.e. the action of μ on f. In such notation, result (3) can be stated as

$$(\mu K) f = \mu(Kf), \tag{4}$$

and this justifies the so-called *measure-kernel-function* notation: $\mu K f$ (without parentheses). When Ω_1 and Ω_2 are finite it is customary to represent a measures μ as a row-vector, a kernel K as a stochastic matrix and a function f as as column vector. In that case, (4) is equivalent to associativity of matrix multiplication.

2.2 Conditional CDFs and PDFs

Here we give a general method for constructing Markov kernels (and via Theorem 1 – joint distributions $\mathbb{P}_{X,Y}$).

Proposition 2. *The following define Markov kernels acting from $(\mathbb{R}, \mathcal{B})$ to itself:*

(a) Let $f_{X|Y}(x|y)$ be a non-negative function jointly measurable in (x, y) and satisfying[1]

$$\int_{\mathbb{R}} f_{X|Y}(x|y)\, dx = 1 \qquad y \in \mathbb{R}, \tag{5}$$

then

$$K(y, dx) = f_{X|Y}(x|y) dx \tag{6}$$

defines a Markov kernel.

[1] Such functions are known as conditional PDFs.

(b) Let $F_{X|Y}(x|y)$ be a function jointly measurable in (x, y), such that $F_{X|Y}(\cdot|y)$ is a CDF [2] for every $y \in \mathbb{R}$. Then there exists a unique Markov kernel s.t.

$$K(y, (a, b]) = F_{X|Y}(b|y) - F_{X|Y}(a|y). \tag{7}$$

Proof. Part (a) is easy: (6) is a measure for every fixed y by (5). The function

$$y \mapsto \int_B f_{X|Y}(x|y)dx$$

is measurable for every $B \in \mathcal{B}$ by Fubini's theorem. For Part (b) again (7) extends to a unique probability measure. We need to verify that the map

$$y \mapsto K(y, B)$$

is measurable for every B. For $B = \bigcup_{i=1}^{n}(a_i, b_i]$ – a finite disjoint union of intervals – this follows from (7) and measurability of finite sums. Then define the collection:

$$\mathcal{L} = \{B \in \mathcal{B} : y \mapsto K(y, B) \text{ –measurable function}\}.$$

We have shown that $\mathcal{L}$ contains algebra of finite unions of intervals $(a, b]$. It is easy to show that $\mathcal{L}$ is a monotone class. Thus, $\mathcal{L} = \mathcal{B}$ and we are done. □

Example. Take PDF f_Y and conditional PDF $f_{X|Y}$. Let

$$\mathbb{P}_Y(dy) = f_Y(y)dy \tag{8}$$
$$K(y, dx) = f_{X|Y}(x|y)dx \tag{9}$$

Then the product measure $\pi = \mathbb{P}_Y \times K$ constructed in Theorem 1 satisfies

$$\pi(dx\, dy) = f_{X|Y}(x|y)f_Y(y)\, dx\, dy$$

In particular, π is a jointly continuous distribution with density $f_{X,Y} = f_{X|Y}f_Y$.

3 DISINTEGRATION OF JOINT DISTRIBUTIONS

Main question we will address here: given $\mathbb{P}_{X,Y}$ does there exist $\mathbb{P}_Y$ and K such that $\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K$?

[2] Such functions $F_{X|Y}$ are known as conditional CDFs.

Definition 2. *Let $\mathbb{P}_{X,Y}$ be a probability measure on $(\mathbb{R}^2, \mathcal{B}^2)$ with marginal $\mathbb{P}_Y$. A Markov kernel $K(y, \cdot)$ is called a regular branch of* ***conditional probability*** *for X given Y (denoted $\mathbb{P}_{X|Y}(\cdot|y)$) if*

$$\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K \tag{10}$$

in the sense of Theorem 1. Equivalently, if

$$\mathbb{P}_{X,Y}[A \times B] = \int_B K(y, A)\, d\mathbb{P}_Y \tag{11}$$

for all $A, B \in \mathcal{B}$. Furthermore, if K is defined via (6) then $f_{X|Y}$ is called a ***conditional PDF****, and if K is defined via (7) then $F_{X|Y}$ is called a* ***conditional CDF****.*

Note: One should not confuse "a regular branch of conditional probability" (which is a Markov kernel) with conditional probability $\mathbb{P}[X \in A|Y]$ (which is a random variable; see below). It should also be clear that neither $\mathbb{P}_{X|Y}$ (a kernel), nor $f_{X|Y}$ (a conditional PDF, when it exists), nor $F_{X|Y}$ (a conditional CDF) are unique. Finally, equivalence of (10) and (11) follows from the fact that $\{A \times B\}$ is a generating p-system for $\mathcal{B} \times \mathcal{B}$.

3.1 Simple case: jointly-continuous $\mathbb{P}_{X,Y}$

For the case of discrete random variables, the conditional CDF is defined by $F_{X|Y}(x \,|\, y) = \mathbb{P}(X \leq x \,|\, Y = y)$, for any y such that $\mathbb{P}(Y = y) > 0$. However, this definition cannot be extended to the continuous case because $\mathbb{P}(Y = y) = 0$, for every y. Instead, we should think of $F_{X|Y}(x \,|\, y)$ as a limit of $\mathbb{P}(X \leq x \,|\, y \leq Y \leq y + \delta)$, as δ decreases to zero. Note that

$$\begin{aligned}
F_{X|Y}(x \,|\, y) &\approx \mathbb{P}(X \leq x \,|\, y \leq Y \leq y + \delta) \\
&= \frac{\mathbb{P}(X \leq x, y \leq Y \leq y + \delta)}{\mathbb{P}(y \leq Y \leq y + \delta)} \\
&\approx \frac{\int_{-\infty}^{x} \int_{y}^{y+\delta} f_{X,Y}(u, v)\, dv\, du}{\delta f_Y(y)} \\
&\approx \frac{\delta \int_{-\infty}^{x} f_{X,Y}(u, y)\, du}{\delta f_Y(y)} \\
&= \frac{\int_{-\infty}^{x} f_{X,Y}(u, y)\, du}{f_Y(y)}.
\end{aligned}$$

This heuristic motivates the next result.

Proposition 3. *Let $f_{X,Y}$ be a joint PDF. Then*

(a) Let f_Y be (any) marginal PDF of Y. Then the following is a conditional CDF of X given Y

$$F_{X|Y}(x \,|\, y) = \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)}\, du,$$

for every y satisfying the following: a) $f_Y(y) > 0$; b) $\int f_{X,Y}(u,y)\, du < \infty$; and c) $\int f_{X,Y}(u,y)\, du = f_Y(y)$. For other y we set $F_{X|Y}(x|y) = 1\{x \geq 0\}$.

(b) The following is a conditional PDF of X given Y

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

for every y such that $f_Y(y) > 0$. For any other y we set $f_{X|Y}(x \,|\, y) = 1\{0 \leq x \leq 1\}$.

Proof. Joint measurability in (x,y) follows from Fubini (in both cases). Next, it is clear that $\mathbb{P}_Y$-measure of all y's satisfying conditions a)-c) is 1. Thus definition of $F_{X|Y}$ for other y's is immaterial for (11). For "good" y's from the DCT we have

$$\lim_{x \searrow x_0} \int_{-\infty}^{x} f_{X,Y}(u,y)\, du = \int_{-\infty}^{x_0} f_{X,Y}(u,y)\, du\,,$$

which shows that $F_{X|Y}(\cdot|y)$ is right-continuous. $F_{X|Y}$ is clearly monotone. The property $\lim_{x\to-\infty} F_{X\,|\,Y}(x \,|\, y) = 0$ follows from the DCT again. Also

$$\lim_{x\to\infty} F_{X|Y}(x \,|\, y) = \int_{-\infty}^{\infty} \frac{f_{X,Y}(u,y)}{f_Y(y)}\, du = 1,$$

since the integral of the numerator is exactly $f_Y(y)$, by condition c).

The proof concludes by a verification of (11) which is left as an exercise. □

3.2 General case: arbitrary $\mathbb{P}_{X,Y}$

Theorem 2 (Disintegration)**.** *Let $\mathbb{P}_{X,Y}$ be a probability measure on $(\mathbb{R}^2, \mathcal{B}^2)$. Then there exists a regular branch of conditional probability $\mathbb{P}_{X|Y}(\cdot|y)$ of X given Y, i.e.*

$$\mathbb{P}_{X,Y} = \mathbb{P}_Y \times \mathbb{P}_{X|Y}.$$

We will prove this result in Section 5.1. We note that similar disintegration works for product spaces other than $\mathbb{R} \times \mathbb{R}$. E.g. X can take values in any complete metric space (not just $\mathbb{R}$), while Y can be arbitrary. For the proof see [Cinlar, Section II.4.2].

4 EXAMPLE: THE BIVARIATE NORMAL DISTRIBUTION

Let us fix some $\rho \in (-1, 1)$ and consider the function, called the **standard bivariate normal PDF**,

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(- \frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \Big).$$

Let X and Y be two jointly continuous random variables, defined on the same probability space, whose joint PDF is f. Therefore, their law satisfies

$$\mathbb{P}_{X,Y}(dx, dy) = f(x, y)\, dx\, dy.$$

Proposition 4. *(a) The function f is a indeed a PDF (integrates to 1).*

(b) The marginal density of X and Y is $N(0, 1)$, the standard normal PDF.

(c) We have $\rho(X, Y) = \rho$. Also, X and Y are independent iff $\rho = 0$.

(d) The conditional density of X, given $Y = y$, is $N(\rho y, 1 - \rho^2)$. In other words,

$$K(y, dx) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}}\, dx$$

is a regular branch of conditional probability for X given Y (i.e. $\mathbb{P}_{X,Y} = N(0, 1) \times K$).

Interpretation of (d): $X = \rho Y + \sqrt{1-\rho^2} Z$, where $Y \perp\!\!\!\perp Z$ are standard normals.

Proof: We will use repeatedly the fact that $1/(\sqrt{2\pi}\sigma) \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ is a PDF (namely, the PDF of the $N(\mu, \sigma^2)$ distribution), and thus integrates to one.

(a)-(b) We note that $x^2 - 2\rho xy + y^2 = x^2 - 2\rho xy + \rho^2 y^2 + (1-\rho^2) y^2$, and

obtain

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{\exp\Big(-\frac{(1-\rho^2)y^2}{2(1-\rho^2)}\Big)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\Big)dx$$
$$= \frac{\exp(-y^2/2)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\Big)dx$$

But we recognize

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\Big)dx$$

as the PDF of the $N(\rho y, 1-\rho^2)$ distribution. Thus, the integral of this density equals one, and we obtain

$$f_Y(y) = \frac{\exp(-y^2/2)}{\sqrt{2\pi}},$$

which is the standard normal PDF. Since $\int_{-\infty}^{\infty} f_Y(y)\,dy = 1$, we conclude that $f(x,y)$ integrates to one, and is a legitimate joint PDF. Furthermore, we have verified that the marginal PDF of Y (and by symmetry, also the marginal PDF of X) is the standard normal PDF, $N(0,1)$.

(c) We have $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY]$, since X and Y are standard normal, and therefore have zero mean. We now have

$$\mathbb{E}[XY] = \iint xyf(x,y)\,dy\,dx.$$

Applying the same trick as above, we obtain for every y,

$$\int xf(x,y)dx = \frac{\exp(-y^2/2)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} x\exp\Big(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\Big)dx.$$

But

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{\infty} x\exp\Big(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\Big)dx = \rho y,$$

since this is the expected value for the $N(\rho y, 1-\rho^2)$ distribution. Thus,

$$\mathbb{E}[XY] = \iint xyf(x,y)\,dx\,dy = \int y\rho y f_Y(y)\,dy = \rho\int y^2 f_Y(y)\,dy = \rho,$$

since the integral is the second moment of the standard normal, which is equal to one. We have established that $\mathrm{Cov}(X,Y) = \rho$. Since the variances of X and Y are equal to unity, we obtain $\rho(X,Y) = \rho$. If X and Y are independent, then $\rho(X,Y) = 0$, implying that $\rho = 0$. Conversely, if $\rho = 0$, then

$$f(x,y) = \frac{1}{2\pi} \exp\Big(- \frac{x^2+y^2}{2} \Big) = f_X(x) f_Y(y),$$

and therefore X and Y are independent. Note that the condition $\rho(X,Y) = 0$ implies independence, for the special case of the bivariate normal, whereas this implication is not always true, for general random variables.

(d) Let us now compute the conditional PDF. Using the expression for $f_Y(y)$

$$\begin{aligned} f_{X|Y}(x \,|\, y) &= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(- \frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)} \Big) \sqrt{2\pi} \exp(y^2/2) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\Big(- \frac{x^2 - 2\rho xy + \rho^2 y^2}{2(1-\rho^2)} \Big) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\Big(- \frac{(x-\rho y)^2}{2(1-\rho^2)} \Big), \end{aligned}$$

which we recognize as the $N(\rho y, 1-\rho^2)$ PDF.

□

We have discussed above the special case of a bivariate normal PDF, in which the means are zero and the variances are equal to one. More generally, the bivariate normal PDF is specified by five parameters, $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$, and is given by

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\Big(- \frac{1}{2} Q(x,y) \Big),$$

where

$$Q(x,y) = \frac{1}{1-\rho^2} \Big[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)}{\sigma_1} \frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \Big].$$

For this case, it can be verified that

$$\mathbb{E}[X] = \mu_1, \quad \mathrm{var}(X) = \sigma_1^2, \quad \mathbb{E}[Y] = \mu_2, \quad \mathrm{var}(Y) = \sigma_2^2, \quad \rho(X,Y) = \rho.$$

These properties can be derived by extending the tedious calculations in the preceding proof.

There is a further generalization to more than two random variables, resulting in the multivariate normal distribution. It will be carried out in a more elegant manner in a later lecture.

5 CONDITIONAL EXPECTATION

Recall that in the discrete case we defined $\mathbb{E}[X|Y=y] = \sum_{x\in\mathbb{R}} x p_{X|Y}(x|y)$. We have also defined $\mathbb{E}[X\,|\,Y]$ to be a a random variable that takes the value $\mathbb{E}[X\,|\,Y=y]$, whenever $Y=y$ and $\mathbb{P}[Y=y]>0$. The general case is more delicate:

Definition 3. *Let X be integrable. A function $g(y)$ is a conditional expectation, denoted $\mathbb{E}[X|Y=y]$, of X given Y if*

$$\mathbb{E}[f(Y)X] = \mathbb{E}[f(Y)g(Y)] \tag{12}$$

for every bounded measurable f. The random variable $g(Y)$, denoted $\mathbb{E}[X|Y]$, is also called a conditional expectation of Y given X. In the special case of $X = 1_B$ we write $\mathbb{P}[B|Y=y]$ or $\mathbb{P}[B|Y]$ to denote a conditional probability of B given Y.

Theorem 3. *Let X, Y be random variables defined on a common probability space and X integrable.*

(a) A conditional expectation $\mathbb{E}[X|Y]$ exists.

(b) If g_1 and g_2 are two conditional expectations of X given Y then

$$\mathbb{P}[g_1(Y) \neq g_2(Y)] = 0\,. \tag{13}$$

(c) If K is a regular branch of conditional probability of X given Y then

$$g(y) = \int_{\mathbb{R}} xK(y, dx) \tag{14}$$

is a conditional expectation $\mathbb{E}[X|Y = y]$. In particular, if a conditional PDF $f_{X|Y}$ exists then

$$g(y) = \int_{\mathbb{R}} x f_{X|Y}(x|y)dx\,.$$

Note: Conditional expectation is not unique. However as (13) shows – this non-uniqueness is immaterial in most cases. Nevertheless, it is a mistake (and a very common one!) to ask for the value of $\mathbb{P}[B|Y = 0]$, which can be set to anything unless $\mathbb{P}[Y = 0] > 0$. The correct question is to find a *function* $y \mapsto \mathbb{P}[B|Y = y]$ (defined upto almost-sure equivalence).

Proof. (a) Let $X = X^+ - X^-$ and define for any Borel set B

$$\nu^+(B) \triangleq \mathbb{E}[1_B(Y)X^+]$$

which evidently defines a finite ($\mathbb{E}[X^+] < \infty$) measure on $(\mathbb{R}, \mathcal{B})$. Furthermore, if $\mathbb{P}_Y(B) = 0$ then $\{Y \in B\}$ has probability 0 and thus $\nu \ll \mathbb{P}_Y$. By Radon-Nikodym theorem there exists a measurable function g^+ such that

$$\mathbb{E}[1_B(Y)X^+] = \mathbb{E}[g^+(Y)1_B(Y)].$$

Similarly, we may define $\nu^-(B)$ via X^- and apply Radon-Nikodym theorem to get g^-. Setting $g = g^+ - g^-$ we have for every Borel set B:

$$\mathbb{E}[1_B(Y)X] = \mathbb{E}[1_B(Y)g(Y)].$$

Thus, $g(Y)$ verifies (12) for all $f = 1_B$. By linearity of expectation (12) is also verified for all simple functions. The general case of bounded f follows by the DCT.

(b) If $g_1(Y)$ and $g_2(Y)$ are two conditional expectations then setting

$$f(Y) = 1\{g_1(Y) > g_2(Y)\} - 1\{g_1(Y) < g_2(Y)\}$$

from (12) we get

$$\mathbb{E}[|g_1(Y) - g_2(Y)|] = 0$$

implying $g_1 = g_2$ with $\mathbb{P}_Y$-probability 1.

(c) By definition if K is a regular branch of conditional expectation then $\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K$, which by Theorem 1 implies

$$\int_{\mathbb{R}^2} \phi(x,y)\mathbb{P}_{X,Y}(dx\,dy) = \int_{\mathbb{R}} \mathbb{P}_Y(dy) \int_{\mathbb{R}} \phi(x,y)K(y,dx)\,.$$

Taking $\phi(x,y) = xf(y)$ and using integrability of X property (12) follows by Fubini. □

Example. One might expect that when X and Y are jointly continuous, then $\mathbb{E}[X \mid Y]$ is a continuous random variable, but this is not the case. To see this, suppose that X and Y are independent, in which case $\mathbb{E}[X \mid Y = y] = \mathbb{E}[X]$, which also implies that $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$. Thus, $\mathbb{E}[X \mid Y]$ takes a constant value, and is therefore a trivial case of a discrete random variable.

Example. We have a stick of unit length $[0,1]$, and break it at X, where X is uniformly distributed on $[0,1]$. Given the value x of X, we let Y be uniformly distributed on $[0,x]$, and let Z be uniformly distributed on $[0,1-x]$. We assume that conditioned on $X = x$, the random variables Y and Z are independent. We are interested in the distribution of Y and Z, their expected values, and the expected value of their product.

It is clear from symmetry that Y and Z have the same marginal distribution, so we focus on Y. Let us first find the joint distribution of Y and X. We have $f_X(x) = 1$, for $x \in [0,1]$, and $f_{Y|X}(y \mid x) = 1/x$, for $y \in [0,x]$. Thus, the joint PDF is

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x)f_X(x) = \frac{1}{x} \cdot 1 = \frac{1}{x}, \qquad 0 \le y \le x \le 1.$$

We can now find the PDF of Y:

$$f_Y(y) = \int_0^1 f_{X,Y}(x,y)\,dx = \int_y^1 f_{X,Y}(x,y)\,dx = \int_y^1 \frac{1}{x}\,dx = \log x \Big|_y^1 = \log(1/y).$$

(check that this indeed integrates to unity). Integrating by parts, we then obtain

$$\mathbb{E}[Y] = \int_0^1 y f_Y(y) dy = \int_0^1 y \log(1/y)\,dy = \frac{1}{4}.$$

The above calculation is more involved than necessary. For a simpler argument, simply observe that $\mathbb{E}[Y \mid X = x] = x/2$, since Y conditioned on $X = x$ is uniform on

$[0, x]$. In particular, $\mathbb{E}[Y \mid X] = X/2$. It follows that $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[X/2] = 1/4$.

For an alternative version of this argument, consider the random variable Y/X. Conditioned on the event $X = x$, this random variable takes values in the range $[0, 1]$, is uniformly distributed on that range, and has mean $1/2$. Thus, the conditional PDF of Y/X is not affected by the value x of X. This implies that Y/X is independent of X, and we have

$$\mathbb{E}[Y] = \mathbb{E}[(Y/X)X] = \mathbb{E}[Y/X] \cdot \mathbb{E}[X] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

To find $\mathbb{E}[YZ]$ we use the fact that, conditional on $X = x$, Y and Z are independent, and obtain

$$\begin{aligned} \mathbb{E}[YZ] = \mathbb{E}\ \mathbb{E}[YZ \mid X] \ &= \mathbb{E}\ \mathbb{E}[Y \mid X] \cdot \mathbb{E}[Z \mid X] \\ = \mathbb{E}\ \frac{X}{2} \cdot \frac{1-X}{2}\Big] &= \int_0^1 \frac{x(1-x)}{4}\, dx = \frac{1}{24}. \end{aligned}$$

Exercise 1. Find the joint PDF of Y and Z. Find the probability $\mathbb{P}(Y + Z \leq 1/3)$. Find $\mathbb{E}[X|Y]$, $\mathbb{E}[X|Z]$, and $\rho(Y, Z)$.

5.1 Other properties of $\mathbb{E}[\cdot|Y]$

We note that many properties of Lebesgue integration carry over without change to conditional expectation:

1. Monotonicity: $X \leq X' \quad \Rightarrow \quad \mathbb{E}[X|Y] \leq \mathbb{E}[X'|Y]$
2. Linearity: $\mathbb{E}[aX + bX'|Y] = a\mathbb{E}[X|Y] + b\mathbb{E}[X'|Y]$
3. MCT: $0 \leq X_n \nearrow X \quad \Rightarrow \quad \mathbb{E}[X_n|Y] \nearrow \mathbb{E}[X|Y]$
4. DCT: $|X_n| \leq Z$, Z-integrable, $X_n \to X$, then $\Rightarrow \mathbb{E}[X_n|Y] \to \mathbb{E}[X|Y]$
5. Fatou's lemma: $X_n \geq 0$, $\mathbb{E}[\liminf_n X_n|Y] \leq \liminf_n \mathbb{E}[X_n|Y]$
6. Jensen's inequality: f convex $\Rightarrow \mathbb{E}[f(X)|Y] \geq f(\mathbb{E}[X|Y])$

Caution: Right-hand sides of each of these implications only hold almost surely!

Proofs of all of these are simple: assume right-hand side is violated on a set E with $\mathbb{P}[Y \in E] > 0$, then using 1_E and (12) construct a counter-example to the uncoditional version of the same property. As an application we prove disintegration theorem:

Proof of Theorem 2. Let $\{r_n\}_{=1}^{\infty}$ be enumeration of rational numbers $\mathbb{Q}$. Denote

$$g_n(y) \triangleq \mathbb{P}[X \in (-\infty; r_n] \,|\, Y = y] \,.$$

By monotonicity property for any k and n such that $r_k \leq r_n$ we have

$$\mathbb{P}[g_k(Y) \leq g_n(Y)] = 1$$

Therefore the set

$$E_0 = \{y : g_k(y) \leq g_n(y) \quad \forall (k,n) : r_k \leq r_n\}$$

has $\mathbb{P}_Y$-measure 1. Similarly, sets

$$E_1 = \{y : \inf_n g_n(y) = 0\} \tag{15}$$

$$E_2 = \{y : \sup_n g_n(y) = 1\} \tag{16}$$

both also have $\mathbb{P}_Y$-measure 1. All together, for every y in the set

$$E \triangleq E_0 \cap E_1 \cap E_2$$

closure of the sequence of points $(r_n, g_n(y))$ on $\mathbb{R} \times [0,1]$ is a graph of a CDF. Thus, we may define:

$$F_{X|Y}(x|y) = \begin{cases} 1\{x \geq 0\}, & y \notin E \\ \sup\{g_n(y) : r_n \leq x\}, & y \in E \end{cases}$$

Notice that

$$y \mapsto F_{X|Y}(x|y)$$

is measurable (as a countable supremum of measurable g_n's). And the function

$$x \mapsto F_{X|Y}(x|y)$$

is right-continuous, monotonically non-decreasing and growing from 0 to 1 on $\mathbb{R}$. Thus by Proposition **??** function $F_{X|Y}$ is jointly measurable. Consequently, it satisfies all requirements of a conditional CDF and by Proposition 2.(b) there exists a Markov kernel $K(y, dx)$ satisfying (7). But then for every set $(-\infty, r_n] \times B$ we have

$$(\mathbb{P}_Y \times K)((-\infty, r_n] \times B) = \int_B \mathbb{P}_Y(dy) K(y, (-\infty, r_n]) \tag{17}$$

$$= \int_B g_n(y) \mathbb{P}_Y(dy) \tag{18}$$

$$= \mathbb{E}[1_{(-\infty, r_n]}(X) 1_B(Y)] \tag{19}$$

$$= \mathbb{P}_{X,Y}((-\infty, r_n] \times B) \,, \tag{20}$$

where first step is by (7), second by definition of $F_{X|Y}(r_n|y)$ and since $\mathbb{P}_Y(E) = 1$, third is by definition of g_n and (12), and fourth is just by definition of $\mathbb{P}_{X,Y}$.

Since sets $(-\infty, r_n] \times B$ form a generating p-system for $\mathcal{B} \times \mathcal{B}$ we conclude

$$\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K$$

which proves the Theorem. □

5.2 Optimality properties of conditional expectations

The conditional expectation $\mathbb{E}[X \mid Y]$ can be viewed as an estimate of X, based on the value of Y. In fact, it is an optimal estimate, in the sense that the mean square of the resulting estimation error, $X - \mathbb{E}[X \mid Y]$, is as small as possible.

Theorem 4. *Suppose that $\mathbb{E}[X^2] < \infty$. Then, for any measurable function $g : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2\big] \leq \mathbb{E}\big[(X - g(Y))^2\big].$$

Proof: We have

$$\begin{aligned}
\mathbb{E}\big[(X - g(Y))^2\big] &= \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2\big] + \mathbb{E}\big[(\mathbb{E}[X \mid Y] - g(Y))^2\big] \\
&\quad +2\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - g(Y))\big] \\
&\geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2\big].
\end{aligned}$$

The inequality above is obtained by noticing that the term $\mathbb{E}\big[(X - g(Y))^2\big]$ is always nonnegative, and that the term $\mathbb{E}[(X - \mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y] - g(Y))$ is of the form $\mathbb{E}[(X - \mathbb{E}[X \mid Y])\psi(Y)$ for $\psi(Y) = \mathbb{E}[X \mid Y] - g(Y)$, and is therefore equal to zero, by Eq. (12). □

Notice that the preceding proof only relies on the property (12). As we have discussed, we can view this as the defining property of conditional expectations, for general random variables. It follows that the preceding theorem is true for all kinds of random variables.

6 MIXED VERSIONS OF BAYES' RULE

Let X be an unobserved random variable, with known CDF, F_X. We observe the value of a related random variable, Y, whose distribution depends on the value of X. This dependence can be captured by a conditional CDF, $F_{Y|X}$. On the basis of the observed value y of Y, would like to make an inference on

the unknown value of X. While sometimes, this inference aims at a numerical estimate for X, the most complete answer, which includes everything that can be said about X, is the conditional distribution of X, given Y. This conditional distribution can be obtained by using an appropriate form of Bayes' rule.

When X and Y are both discrete, Bayes' rule takes the simple form

$$p_{X|Y}(x \mid y) = \frac{p_X(x)p_{Y|X}(y \mid x)}{p_Y(y)} = \frac{p_X(x)p_{Y|X}(y \mid x)}{\sum_{x'} p_X(x')p_{Y|X}(y \mid x')}.$$

When X and Y are both continuous, Bayes' rule takes a similar form,

$$f_{X|Y}(x \mid y) = \frac{f_X(x)f_{Y|X}(y \mid x)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y \mid x)}{\int f_X(x')f_{Y|X}(y \mid x')\,dx'},$$

which follows readily from the definition of the conditional PDF.

It remains to consider the case where one random variable is discrete and the other continuous. Suppose that K is a discrete random variable and Z is a continuous random variable. We describe their joint distribution in terms of a function $f_{K,Z}(k, z)$ that satisfies

$$\mathbb{P}(K = k, Z \leq z) = \int_{-\infty}^{z} f_{K,Z}(k, t)\,dt.$$

We then have

$$p_K(k) = \mathbb{P}(K = k) = \int_{-\infty}^{\infty} f_{K,Z}(k, t)\,dt,$$

and[3]

$$F_Z(z) = \mathbb{P}(Z \leq z) = \sum_k \int_{-\infty}^{z} f_{K,Z}(k, t)\,dz = \int_{-\infty}^{z} \sum_k f_{K,Z}(k, t)\,dz,$$

which implies that

$$f_Z(z) = \sum_k f_{K,Z}(k, z),$$

is the PDF of Z.

Note that if $\mathbb{P}(K = k) > 0$, then

$$\mathbb{P}(Z \leq z \mid K = k) = \int_{-\infty}^{z} \frac{f_{K,Z}(k, t)}{p_K(k)}\,dt,$$

[3]The interchange of the summation and the integration can be rigorously justified, because the terms inside are nonnegative.

and therefore, it is reasonable to define

$$f_{Z|K}(z \mid k) = f_{K,Z}(k, z)/p_K(k).$$

Finally, for z such that $f_Z(z) > 0$, we define $p_{K|Z}(k \mid z) = f_{K,Z}(k, z)/f_Z(z)$, and interpret it as the conditional probability of the event $K = k$, given that $Z = z$. (Note that we are conditioning on a zero probability event; a more accurate interpretation is obtained by conditioning on the event $z \leq Z \leq z + \delta$, and let $\delta \to 0$.) With these definitions, we have

$$f_{K,Z}(k, z) = p_K(k) f_{Z|K}(z \mid k) = f_Z(z) p_{K|Z}(k \mid z),$$

for every (k, z) for which $f_{K,Z}(k, z) > 0$. By rearranging, we obtain two more versions of the Bayes' rule:

$$f_{Z|K}(z \mid k) = \frac{f_Z(z) p_{K|Z}(k \mid z)}{p_K(k)} = \frac{f_Z(z) p_{K|Z}(k \mid z)}{\int f_Z(z') p_{K|Z}(k \mid z')\, dz'},$$

and

$$p_{K|Z}(k \mid z) = \frac{p_K(k) f_{Z|K}(z \mid k)}{f_Z(z)} = \frac{p_K(k) f_{Z|k}(z \mid k)}{\sum_{k'} p_K(k') f_{Z|K}(z \mid k')}.$$

Note that all four versions of Bayes' rule take the exact same form; the only difference is that we use PMFs and summations for discrete random variables, as opposed to PDFs and integrals for continuous random variables.

6.436J / 15.085J Fundamentals of Probability

Fall 2018