A review of basic probability theory (Cont)

Math 485

December 3, 2014

1 Conditional expectation

1.1 Conditional distribution, conditional density

We have discussed conditional probability $P(A|B)$, which is the probability that A happened given the knowledge that B has happened. In a similar way, for 2 RVs X, Y, we can talk about the probability that X takes some value x given that we know Y has taken some y. If X and Y are correlated in some way, the fact that we have seen Y taking some value should change the probability that X taking value x. Formally, we define, for 2 discrete RVs X, Y

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

For continuous RVs, we cannot talk about the probablity that X takes some value, given that we have observed Y taking some value. The reason is the probability that Y taking some value is 0, since it is a continuous RV. This poses a slight problem, since in reality, we always observe Y taking some particular value, even if it is a continuous RV (think about the amount of time you wait for the bus to arrive, for example. You always have to wait a particular amount of time until the bus arrives, even if the probability that the continuous random variable representing the time you wait taking that particular value is 0). So for continuous RVs, we talk about the conditional density instead. Formally, we define, for 2 continuous RVs X, Y

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

Remark: In the two formulas above, we think of y as *fixed*, and x as taking any possible values in the range of X. Thus the conditional distribution, or conditional density, is a function of x, given a fixed value y. Moreover, for a fixed y, the conditional distribution (or probability density), is a probability distribution (or density). That is

$$\sum_x P(X = x|Y = y) = 1;$$

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1$$

.

Exercise: Prove these two equalites.

1.2 Conditional probability and conditional expectation

1.2.1 Discrete

Let X, Y be discrete RVs. The conditional probability $P(X = i|Y = j)$ was defined naturally using the definition of conditional expectation as above. Note that we also have

$$P(X \leq k|Y = j) = \sum_{i \leq k} P(X = i|Y = j).$$

In this way, for every y, conditioned on $Y = y$, $P(X < a|Y = y)$ is a proper cumulative distribution function, even though $P(Y = y) = 0$. This is related to the notion of regular conditional probability distribution, discussed below.

We define the conditional expectation of X, given $Y = y$ as

$$E(X|Y = y) = \sum_x xP(X = x|Y = y).$$

1.2.2 Continuous

Let X, Y be continuous RVs. Note that we can NOT define $P(X < a|Y = y)$ using the definition of conditional expectation, because $P(Y = y) = 0$. However, we can define it as followed:

$$P(X < a|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y)dx.$$

We define the conditional expectation of X, given $Y = y$ as

$$E(X|Y = y) = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx.$$

Interpretation: Besides the fact that conditional expectation is the average (or mean) value of X given $Y = y$, it is also the *best guess* of X given $Y = y$, in some precise sense that we will discuss below.

Remark: Note that in these definitions, $E(X|Y = y)$ is *a real number*. This will be constrasted with $E(X|Y)$, which is a RV, the definition of which is given below.

1.3 Abstract definition of conditional expectation

1.3.1 Motivation

The above definitions of $E(X|Y = y)$, while useful, is rather restrictive. It is because we do not have to observe the value of Y to be able to talk about the expectation of X conditioned on Y in a meaningful way. An example will explain. It is clear that the stock price of today depends on the stock price of yesterday (for simplicity let's suppose that stock price only changes discretely from day n to day $n + 1$). Suppose we are at day 0, which is today, and we want to discuss our "expectation", or our best guess, of the stock price on day $n + 1$, the guess being made on day n. It is clear that on day n, we have the knowledge of the stock price of that day, say S_n. So what we're asking for is $E(S_{n+1}|S_n)$. Since we are still at day 0, we do not know what value S_n is, it is a RV to us. However, to discuss our action on day n, in anticipation of day $n + 1$, it is necessary that we make sense of the notion $E(S_{n+1}|S_n)$. Thus we need an abstract definition of conditional expectation, one that doesn't require us to plug in an observed value for the RV being conditioned on. We will als refer to this as *the measure theoretic definition* of conditional expectation.

1.3.2 Definition

Definition 1.1. *Let X, Y be RVs. The conditional expectation $E(X|Y)$ is a function of Y, such that for any function g, we have*

$$E\big[E(X|Y)g(Y)\big] = E\big[Xg(Y)\big].$$

Remark: Note that in contrast with the above, as we already said, $E(X|Y)$ is a RV, since it is a function of Y (in some trivial case it could be the constant function, but this does not happen usually). The interpretation of the equality in the definition is that as far as taking expectation with respect to function of Y, it does not matter if we use the conditional expectation $E(X|Y)$ or X itself. Thus the conditional expectation $E(X|Y)$ is a guess of X, in terms of the random variable Y, which satisfies some "indifference" property in terms of expectation.

Perhaps a more satisfactory property of $E(X|Y)$ is that not only it is a guess of X given Y, it is *the best* guess of X given Y in the following sense:

Lemma 1.2. *Let X, Y be RVs. Then for any function g we have*

$$E\Big([E(X|Y) - X]^2\Big) \le E\Big([g(Y) - X]^2\Big).$$

Proof. See homework.

1.3.3 Some elementary properties

The definition (1.1) unfortunately does not, most of the time, give us an easy way to compute what $E(X|Y)$ is. So the followings are some elementary properties of conditional expectation that will help us do that. You should try to prove these properties yourself.

a. $E(E(X|Y)) = E(X)$.

b. $E(aX + bY|Z) = a(EX|Z) + bE(Y|Z), a, b$ constant .

c. If X is independent of Y then

$$E(X|Y) = E(X).$$

d. For any function g,

$$E(g(Y)X|Y) = g(Y)E(X|Y).$$

e. *The independence lemma:* If X is independent of Y then for any function g

$$E\big[g(X,Y)|Y\big] = E\big[g(X,y)\big]|_{y=Y}.$$

(Properties f and g are not used in this course except in the discussion of regular conditional probability. You can skip these.)

f. If $X_n \ge 0, X_n \uparrow X$ then $E(X_n|\mathcal{G}) \uparrow E(X|\mathcal{G})$.

g. If $X \in L_2(\Omega, \mathcal{F}, P)$ then $E(X|\mathcal{G})$ is the orthogonal projection of X onto the subspace $L_2(\Omega, \mathcal{G}, P)$ in the Hilbert space $L_2(\Omega, \mathcal{F}, P)$ with inner product $\langle X, Y\rangle := E(XY)$.

Remark: The expression $E\big[g(X,y)\big]|_{y=Y}$ means that we just evaluate $E\big[g(X,y)\big]$ as a *regular expectation* (it is only a random variable in terms of X, y is understood to be a constant (or just a dummy variable) here. Note that $E\big[g(X,y)\big]$ is *a function of* y. Thus we are free to plug in the random variable Y after we compute what $E\big[g(X,y)\big]$ is.

Example 1.3. *Let X be a Bernoulli(1/2) random variable and Y has Normal(0,1) distribution, X independent of Y. Compute $E(Y^X|Y)$.*

Ans: We have

$$E(y^X) = y^0 \cdot \frac{1}{2} + y^1 \cdot \frac{1}{2} = \frac{1}{2}(1+y).$$

Thus y the independence lemma, $E(Y^X|Y) = \frac{1}{2}(1+Y)$. Note how the distribution of Y is irrelevant in this computation.

1.4 Expectation conditional on more than one random variables

In applications, a random variable X may be correlated to not just 1 random variable Y, but possibly to n random variables $Y_1, Y_2, \cdots, Y_n$ (it is reasonable to build a model of stock so that the stock price today does not just depend on its performance yesterday, but on its performance in the past month). To discuss the behavior of X given our observations of $Y_1, \cdots Y_n$, we need to extend our notion of conditional expectation to more than 1 random variable. The extension actually is straightforward.

Definition 1.4. *Let $X, Y_1, Y_2, \cdots, Y_n$ be RVs. The conditional expectation $E(X\big|Y_1, Y_2, \cdots, Y_n)$ is a function of $Y_1, Y_2, \cdots, Y_n$, such that for any function g, we have*

$$E\big[E(X\big|Y_1, Y_2, \cdots, Y_n)g(Y_1, Y_2, \cdots, Y_n)\big] = E\big[Xg(Y_1, Y_2, \cdots, Y_n)\big].$$

Remark: Actually in section (1.3), there is no restriction on what the RV Y can be. Thus one could select it to be a *multi-dimensional RV*, effectively making it a random vector with n components

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}.$$

Even more generaly, X itself can also be a multi-dimensional RV,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_m \end{bmatrix}.$$

Thus we see that we have covered the case of expectation conditional on more than one random variables:

$$E(X_1, X_2, \cdots X_m | Y_1, Y_2, \cdots, Y_n)$$

in section (1.3), including the elementary properties. One just needs to interpret the symbol accordingly, for example in $E(aX + bY|Z)$, a, b has to be understood as vector, aX and bY as vector dot products if X, Y are multi-dimensional RV.

1.5 Probability as an expectation

You may observe that in the abstract definition of conditional expectation, we did not mention about conditional probability. Surely we would want to have a definition for $P(X \leq x|Y)$. It turns out that our definition of conditional expectation *already covers conditional probability as a special case.* To be precise, we first need to introduct the following so-called indicator function of an event E, denoted as $\mathbf{1}_E$

$$\begin{aligned} \mathbf{1}_E(\omega) &= 1 \text{ if } \omega \in E \\ &= 0 \text{ if } \omega \notin E. \end{aligned}$$

Basically the indicator function is a logical indicator, it's 1 if E happens and 0 if E does not happen. For example $\mathbf{1}_{\{0<1\}} = 1$ and $\mathbf{1}_{\{1+1<3\}} = 0$. But now note that suppose we have a random variable X, and say it has a density function $f_X(x)$ then

$$\begin{aligned} E(\mathbf{1}_{\{X \leq x\}}) &= \int_{-\infty}^{\infty} \mathbf{1}_{\{y \leq x\}}(y) f_X(y) dy \\ &= \int_{-\infty}^{x} f_X(y) dy = P(X \leq x). \end{aligned}$$

where the second equality is because $\mathbf{1}_{\{y \leq x\}} = 0$ for all values of $y > x$ so we just stop the integration limit at x. Similarly, you can check that

$$\begin{aligned} E(\mathbf{1}_{\{X \geq x\}}) &= P(X \geq x) \\ E(\mathbf{1}_{\{X = x\}}) &= P(X = x). \end{aligned}$$

Thus probability can be expressed as an expectation. More importantly for us, this is still true at the conditional expectation level. More precisely we have the following

Lemma 1.5. *Let X, Y be random variables. Let $f(Y) = E(\mathbf{1}_{\{X \le x\}}|Y)$. Then $f(y) = P(X \le x|Y = y)$ where $P(X \le x|Y = y)$ is understood in the sense of section* (1.1). *Similarly for $P(X \ge x|Y = y), P(X = x|Y = y)$.*

Remark: For a fixed x, the expression $\mathbf{1}_{\{X \le x\}}$ here is understood as function of X. Thus the expression $E(\mathbf{1}_{\{X \le x\}}|Y)$ is understood in the sense of $E(g(X)|Y)$ where $g(X)$ is just a random variable.

2 Connection between the measure theoretic and classical definition of conditional expectations

2.1 Discrete RVs

Let X, Y be two RVs. We have seen that we can define $E(X|Y)$ abstractly via definition (1.1). Suppose X, Y are both discrete. Then we also have alternative definitions of $E(X|Y = y)$ via classical probability theory. How are these two connected?

Note that $E(X|Y)$ by definition if a function of Y. Thus we can write $E(X|Y) = g(Y)$ for some function g. On the other hand, $E(X|Y = y)$ is also clearly a function of y. So you can expect that $\forall y$ on the event $\{Y = y\}$

$$E(X|Y) = E(X|Y = y).$$

That is

$$E(X|Y)\mathbf{1}_{Y=y} = E(X|Y = y)\mathbf{1}_{Y=y}.$$

Proof. We need to check that for any function $g(Y)$

$$E\Big[E(X|Y = y)\mathbf{1}_{Y=y}g(Y)\Big] = E\Big[X\mathbf{1}_{Y=y}g(Y)\Big].$$

The LHS is equal to

$$\begin{aligned} E(X|Y = y)g(y)P(Y = y) &= \sum_i \frac{iP(X = i, Y = y)}{P(Y = y)} g(y)P(Y = y) \\ &= \sum_i iP(X = i, Y = y)g(y) \\ &= E\Big[Xg(y)\mathbf{1}_{Y=y}\Big] = RHS. \end{aligned}$$

2.2 Continuous RVs

What about the case when both X, Y are continuous? Here we cannot use the above criterion, as the event $\{Y = y\}$ has probability 0. Rather we will turn it around, and observe that since $E(X|Y = y)$ is a function of y, we can also write $E(X|Y = y) = g(y)$. So now we can plug the RV Y into the function g and we claim

$$E(X|Y) = g(Y), P \text{ a.s.},$$

where the a.s. notation means the equality holds outside an event of probability 0 with respect to P.

Proof.

$$E(X|Y = y) = \int_{-\infty}^{\infty} x \frac{f_{XY}(x, y)}{f_Y(y)} dx.$$

Therefore

$$g(Y) = \int_{-\infty}^{\infty} x \frac{f_{XY}(x, Y)}{f_Y(Y)} dx$$

We need to check that for any function $h(Y)$

$$E\Big(h(Y) \int_{-\infty}^{\infty} x \frac{f_{XY}(x, Y)}{f_Y(Y)} dx\Big) = E\Big(Xh(Y)\Big).$$

But the LHS is equal to

$$\begin{aligned} & \int_{-\infty}^{\infty} h(y) \Big[\int_{-\infty}^{\infty} x \frac{f_{XY}(x, y)}{f_Y(y)} dx \Big] f_Y(y) dy \\ = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y) x f_{XY}(x, y) dx dy = E(Xh(Y)) = RHS. \end{aligned}$$

3 Law of large number

3.1 The theorem

Theorem 3.1. *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed (abbreviated as i.i.d.) RVs such that $E|X_1| < \infty$. Then with probability 1,*

$$\frac{\sum_{i=1}^{n} X_k}{n} \to E(X_1).$$

Notation: It is usually denoted that $S_n = \sum_{i=1}^{n} X_k$, thus one usually sees $\frac{S_n}{n} \to E(X_1)$ in the statement of the law of large number (LLN).

Interpretation: Suppose that you play a game where your winning is random, which is represented by a RV X. As you play this game many times, you will find that your average earning (over time) is approximately the expected value of X.

3.2 Application

We'll give one application of the LLN, in pricing of a random game: Suppose there is a game of tossing a fair coin, where if the coin turns up H then you get paid 3 dollars. If it turns up T, then you get paid 1 dollars. Question: What is the fair price to charge for this game?

Ans: The fair price to charge for this game is $3\frac{1}{2} + 1\frac{1}{2} = 2$ (dollars). But can you explain why this is the fair price? The reason is the LLN. If this game is played *only once* (an important point, which we'll come back later when we discuss the fair price of financial instrument) then it is not clear that the price is fair. However, the assumption here is that the game will be played *many times,* by potentially many different players. Thus each player's winning is an independent, identically distributed random variable, which takes values 3 and 1 with probability 1/2 each. The total amount of money the house has to pay to these players, after n games have been played, is $\sum_{i=1}^{n} X_i$. By the LLN, this is approximately $nE(X_1)$, which is $2n$, which is the total amount charged by the house. So the house comes out even and this is a fair price for the game.

Remark: This is the main principle behind casino's operation (and profitability). Of course the players are not charged to play the games in the casino. But the game is set up so that the expectation is negative (even if you bet on a roulette table, say on an even number, your chance of winning is still less than 1/2, since there is a 0 and double 0's). Thus by the LLN, with a lot of customers, the casino will have a positive profit. Note that the LLN does allow for an occasional incident where someone plays 1 single game and win big. But if you play a lot of games at the casino, the LLN says that you will lose money eventually.

4 Central limit theorem

4.1 The theorem

The LLN gives us an estimate of $\frac{S_n}{n}$ (it is approximately $E(X_1)$). However, for various reasons, we may want a more precise estimate than that. Note that the LLN says nothing about how close to $E(X_1)$ $\frac{S_n}{n}$ is, or (perhaps surprisingly) what distribution we may approximate $\frac{S_n}{n}$ with. It is surprising because we do not have any restriction on the distribution of each individual X_i, but it turns out that the approximate distribution of $\frac{S_n}{n}$ is the normal distribution. The precise statement is as followed:

Theorem 4.1. *Let $X_1, X_2, \cdots$ be i.i.d. RV such that $E|X_1|^2 < \infty$. We will also denote $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Then for any real number x,*

$$P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x) \to P(Z \leq x),$$

where Z has standard normal $(N(0,1))$ distribution.

4.2 Application

The central limit theorem is used to estimate probability of the *sum* or the *average* of an i.i.d. sequence of RVs. Determining which case to use requires a close reading into the problem.

Example 4.2. *The bus arrives at the Hill center according to a uniform $[0, 12]$ distribution. Suppose you wait at the Hill center bus stop for 30 days. What is the approximate probability that your* **average** *wait time is more than 5 minutes?*

Ans: Let $X_1, X_2, \cdots$ be i.i.d. $U[0, 12]$. Then $E(X_1) = 6$ and $Var(X_1) = 12$. Thus

$$P(\frac{S_{30}}{30} \geq 5) = P(\frac{S_{30} - 6 \times 30}{\sqrt{12 \times 30}} \geq \frac{5\sqrt{30}}{\sqrt{12}} - \frac{6\sqrt{30}}{\sqrt{12}}) \approx P(Z \geq -1.58).$$

Example 4.3. *The earning per day of a casino is distributed as an Exponential(1) RV. (1 here stands for 1 million, we omit the unit). What is the approximate probability that the casino's earning in 1 month is more than 35 millions?*

Ans: Note that here we're asked for the **total** *earning. Thus let $X_1, X_2, \cdots$ be i.i.d. $Exp(1)$. Then $E(X_1) = 1$ and $Var(X_1) = 1$. Thus*

$$P(S_{30} \geq 35) = P(\frac{S_{30} - 30}{\sqrt{30}} \geq \frac{5}{\sqrt{30}}) \approx P(Z \geq \frac{5}{\sqrt{30}}).$$

5 Appendix - On regular conditional probability distribution

(This section is not required for the course. You can skip this.)

In the above discussion, you see that we can define, for X, Y jointly continuous

$$P(X < a|Y = y) := \int_{-\infty}^{a} f_{X|Y}(x|y)dx,$$

even though $P(Y = y) = 0$. This is connected with the notion of regular conditional probability distribution (r.c.p.d). We will discuss r.c.p.d. and the connection in this section.

5.1 R.c.p.d

Let $X : (\Omega, \mathcal{F}, P) \to (S, \mathcal{S})$ be a RV taking values in a general space S with sigma-field $\mathcal{S}$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub sigma-field.

For any set $A \in \mathcal{S}$, the conditional probability of $X \in A$ given $\mathcal{G}$ is given by the conditional expectation

$$P(\omega, A) := E(\mathbf{1}_A(X)|\mathcal{G}). \tag{1}$$

As we vary A in $\mathcal{S}$, we obtain a map from $\Omega \times \mathcal{S}$ to $[0, 1]$.

The question is: when we consider simultaneuously all $A \in \mathcal{S}$, that is when we fix a particular ω, can we ensure that for P almost every ω, $P(\omega, \cdot)$ is a probability measure (on $\mathcal{S}$)? The answer is non trivial.

First note that for each A, $P(\cdot, A)$ is only defined almostsuredly via (1). That is, we are free to modify $P(\cdot, A)$ outside a set of probability 0.

Thus, in particular, for any countable collection of disjoint sets $A_n \in \mathcal{S}$, we can define $P(\omega, A_n)$ and $P(\omega, \cup_n A_n)$ such that

$$P(\omega, \cup_n A_n) = \sum_n P(\omega, A_n)$$

holds outside an exceptional set $N^{\{A_n\}}$ of probability 0. The notation $N^{\{A_n\}}$ is used to emphasize the dependence of the exceptional set on the sequence $\{A_n\}$.

We desire a version of $P(\cdot, \cdot)$ (that is a definition of $P(\omega, A)$ for any $A \in \mathcal{S}$ outside a set $N \subseteq \Omega$ of probability 0. This set N is defined **independent of** all the sets

A), such that $P(\omega, \cdot)$ satisfies the above countable additivity property for **any** infinite collection of disjoint sets.

Since there are possibly uncountable number of such collections of sets, the corresponding exceptional sets $N^{\{A_n\}}$ of probability 0 could add up to a set with positive probability, or even become non-measurabale.

Definition 5.1. *Let $(\Omega, \mathcal{F}, P), \mathcal{G}$ and $X : (\Omega, \mathcal{F}, P) \to (S, \mathcal{S})$ be as above. A family of probability distributions on $(S, \mathcal{S})$, denoted by $P(\omega, \cdot)_{\omega \in \Omega}$ is called a r.c.d. of X given $\mathcal{G}$ if for each $A \in \mathcal{S}$, $P(\cdot, A) = E(\mathbf{1}_A(X)|\mathcal{G})$-a.s. . When $(S, \mathcal{S}) = (\Omega, \mathcal{F})$ and $X(\omega) = \omega$, $P(\omega, \cdot), \omega \in \Omega$ is called a r.c.p on $\mathcal{F}$ given $\mathcal{G}$.*

Remark: The above definition captures the two desirable properties of a r.c.p.d that we discussed above: For each $\omega, P(\omega, \cdot)$ is a probability on $\mathcal{S}$. Second, it agrees with the conditional expectation given $\mathcal{G}$ almost surely. Note that the family is given *a priori* to the discussion about its agreement with the conditional expectation. In this way it is defined for all $\omega \in O$. The countable additivity property follows from the definition because

$$
\begin{aligned}
P(\cdot, A_n) &= E(\mathbf{1}_{A_n}(X)|\mathcal{G})\text{-a.s.} \\
P(\cdot, \cup_n A_n) &= E(\mathbf{1}_{\cup_n A_n}(X)|\mathcal{G})\text{-a.s.}
\end{aligned}
$$

Thus we can choose an exepctional set $N^{\{A_n\}}$ of probability 0 such that outside of $N^{\{A_n\}}$

$$
P(\cdot, \cup_n A_n) = E(\mathbf{1}_{\cup_n A_n}(X)|\mathcal{G}) = \sum_n E(\mathbf{1}_{A_n}(X)|\mathcal{G}) = \sum_n P(\cdot, A_n),
$$

where the second equality follows from the property that if $X_n \geq 0, X_n \uparrow X$ then $E(X_n|\mathcal{G}) \uparrow E(X|\mathcal{G})$.

If X has a r.c.d given $\mathcal{G}$ then conditional expectations of X given $\mathcal{G}$ can be expressed as integrals over the r.c.d.

Proposition 5.2. *Let $P(\omega, \cdot)_{\omega \in \Omega}$ be a r.c.d of X given $\mathcal{G}$. Then for any Borel measurable funcion $f : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B})$ with $E(f(X)) < \infty$, we have*

$$
E(f(X)|\mathcal{G})(\omega) = \int f(x) P(\omega, dx), \text{ for } P \text{ a.e. } \omega.
$$

5.1.1 Examples of r.c.p.d., discrete RV

Let X, Y have discrete distribution. In this case WLOG we can consider $\Omega = \mathbb{Z}^2, X, Y$ are coordinate mappings: for $\omega = (\omega_1, \omega_2)$:

$$\begin{aligned} X(\omega) &= \omega_1 \\ Y(\omega) &= \omega_2. \end{aligned}$$

We let $\mathcal{F} = \mathcal{F}^{X,Y}, \mathcal{G} = \mathcal{F}^Y$. We define a probability measure on $\mathcal{F}$ simply by

$$P(\omega) = P(X = \omega_1, Y = \omega_2).$$

(In this way, if ω is such that X cannot take value ω_1 or Y cannot take value ω_2, we simply assign $P(\omega) = 0$).

For any $\omega \in \Omega$ and any $A \in \mathcal{B}(\mathbb{R})$ we define

$$\begin{aligned} P(\omega, A) &:= P(X \in A|\mathcal{G})(\omega) := \sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} \frac{P(X = \widetilde{\omega}_1, Y = \omega_2)}{P(Y = \omega_2)} \\ &= \frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(X = \widetilde{\omega}_1, Y = \omega_2)}{P(Y = \omega_2)} \\ &= \frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}{\sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega})} \\ &= P(X \in A|Y = \omega_2). \end{aligned}$$

We claim that $P(\omega, A)$ is the r.c.d. of X given $\mathcal{G}$.

First, for any fixed ω, it is clear that $P(\omega, A)$ is a probability measure. To check that $P(\omega, A) = E(\mathbf{1}_A(X)|\mathcal{G})(\omega)$, for every ω, we need to check that for any function $g(Y)$;

$$\int P(\omega, A) g(Y)(\omega) dP(\omega) = \int \mathbf{1}_A(X)(\omega) g(Y)(\omega) dP(\omega).$$

The LHS is equal to

$$\sum_{\omega} \frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}{\sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}(\omega) g(Y)(\omega) P(\omega).$$

Observe that for any ω, $\frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}{\sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}(\omega)$ and $g(Y)(\omega)$ only depend on ω_2. That is, if ω and ω' are such that $\omega_2 = \omega_2'$ then

$$\frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}{\sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}(\omega) = \frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}{\sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}(\omega')$$

and $g(Y)(\omega) = g(Y)(\omega')$. Therefore the LHS is actually equal to

$$
\begin{aligned}
& \sum_{\omega} \left(\frac{\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega})}{\sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega})} (\omega) g(Y)(\omega) \right) \sum_{\tilde{\omega}:\tilde{\omega}_2 = \omega_2} P(\tilde{\omega}) \\
= & \sum_{\omega} \left(\sum_{\tilde{\omega}:\tilde{\omega}_1 \in A, \tilde{\omega}_2 = \omega_2} P(\tilde{\omega}) \right) (\omega) g(Y)(\omega) \\
= & \int \mathbf{1}_A(X)(\omega) g(Y)(\omega) dP(\omega).
\end{aligned}
$$

5.1.2 Examples of r.c.p.d., continuous RV

Let X, Y have continuous distribution. In this case we consider $\Omega = \mathbb{R}^2, X, Y$ are coordinate mappings: for $\omega = (\omega_1, \omega_2)$:

$$
\begin{aligned}
X(\omega) &= \omega_1 \\
Y(\omega) &= \omega_2.
\end{aligned}
$$

We let $\mathcal{F} = \mathcal{F}^{X,Y}, \mathcal{G} = \mathcal{F}^Y$. For a set $E \in \mathcal{F}$ we define

$$
P(E) = \int_E f_{XY}(x, y) dx dy.
$$

(Again, in this way if there is a set $E \in \mathcal{B}(\mathbb{R}^2)$ such that X, Y cannot take values in E, we simply assign $P(E) = 0$).

For any $\omega \in \Omega$ and any $A \in \mathcal{B}(\mathbb{R})$ we define

$$
\begin{aligned}
P(\omega, A) &:= P(X \in A | \mathcal{G})(\omega) := \frac{\int_A f_{XY}(x, \omega_2) dx}{f_Y(\omega_2)} \\
&= \frac{\int_A f_{XY}(x, \omega_2) dx}{\int_{-\infty}^{\infty} f_{XY}(x, \omega_2) dx} = P(X \in A | Y = \omega_2).
\end{aligned}
$$

We claim that $P(\omega, A)$ is the r.c.d. of X given $\mathcal{G}$.

First, for any fixed ω, it is clear that $P(\omega, A)$ is a probability measure. To check that $P(\omega, A) = E(\mathbf{1}_A(X)|\mathcal{G})(\omega)$, for every ω, we need to check that for any function $g(Y)$;

$$
\int P(\omega, A) g(Y)(\omega) dP(\omega) = \int \mathbf{1}_A(X)(\omega) g(Y)(\omega) dP(\omega).
$$

The LHS is equal to (in the following $\omega = (x, y) = (\omega_1, \omega_2)$)

$$
\int_{\mathbb{R}^2} P(\omega, A) g(Y)(\omega) f_{XY}(\omega) dx dy = \int_{\mathbb{R}^2} \frac{\int_A f_{XY}(x, \omega_2) dx}{\int_{-\infty}^{\infty} f_{XY}(x, \omega_2) dx} (\omega) g(Y)(\omega) f_{XY}(\omega) dx dy.
$$

Again note that both $\frac{\int_A f_{XY}(x,\omega_2)dx}{\int_{-\infty}^{\infty} f_{XY}(x,\omega_2)dx}(\omega)$ and $g(Y)(\omega)$ only depend on ω_2. Thus the LHS is equal to

$$\begin{aligned}
& \int_{\mathbb{R}} \left(\frac{\int_A f_{XY}(x,\omega_2)dx}{\int_{-\infty}^{\infty} f_{XY}(x,\omega_2)dx}(\omega) g(Y)(\omega) \right) \left(\int_{\mathbb{R}} f_{XY}(\omega)dx \right) dy \\
= & \int_{\mathbb{R}} \left(\int_A f_{XY}(x,\omega_2)dx \right) (\omega) g(Y)(\omega) dy \\
= & \int_{\mathbb{R}} \int_A f_{XY}(x,y) g(Y) dx dy = \int \mathbf{1}_A(X)(\omega) g(Y)(\omega) dP(\omega).
\end{aligned}$$

5.1.3 Existence of r.cp.d.

It is clear from our above discussion that we do not always have a family of r.c.p.d. The following gives a sufficient condition for the existence of r.c.p.d.

Theorem 5.3. *Let $X : (\Omega, \mathcal{F}, P) \to (S, \mathcal{S})$ be a RV taking values in a general space S with sigma-field $\mathcal{S}$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub sigma-field. If S is a* **complete separable metric space** *with Borel sigma field $\mathcal{S}$ then there exists a r.c.p.d. family $P(\omega, \cdot)_{\omega \in \Omega}$ of X given $\mathcal{G}$.*

Remark: From our examples, in the discrete case it is clear that the r.c.p.d. always exists. The key there is because X, Y takes values on a space with *countably many* values, so that we can re-cast the probability space as $\mathbb{Z}^2$. This demonstrates the condition requiring S being a separable metric space

In the jointly continuous case the r.c.p.d also exists, but it is not clear how the separability comes into play, as we used the density to define the r.c.p.d. there. One can suspect that the existence of the density f_{XY}, or the Radon-Nikodym derivative of the distribution of X, Y with respect to the Lebesgue measure on $\mathbb{R}^2$, has to do with the separability of $\mathbb{R}^2$.