Lecture 5: Conditional Expectation

Tianpei Xie

Aug.4th., 2015

Contents

1 Recall Signed Measure and Lebesgue Decomposition

1.1 Signed Measure

- **Definition** (***Signed Measure***)
Let $(X, \mathscr{B})$ be a measure space. A ***signed measure*** on $(X, \mathscr{B})$ is a function $\nu : \mathscr{B} \to [-\infty, +\infty]$ such that

  1. (***Emptyset***) $\nu(\emptyset) = 0$;
  2. (***Finiteness in One Direction***) ν assumes at most one of the values $\pm\infty$;
  3. (***Countable Additivity***) if $\{E_j\}$ is a sequence of disjoint sets in $\mathscr{B}$, then $\nu\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \nu(E_j)$, where the latter converges absolutely if $\nu\left(\bigcup_{j=1}^{\infty} E_j\right)$ is finite.

- **Definition** A measure μ is ***finite***, if $\mu(X) < \infty$; μ is σ***-finite***, if $X = \bigcup_{k=1}^{\infty} U_k$, $\mu(U_k) < \infty$.

- **Definition** (***Positive Measure***)
If ν is a signed measure on $(X, \mathscr{B})$, a ***set*** $E \in \mathscr{B}$ is called ***positive*** (resp. ***negative***, ***null***) for ν if $\nu(F) \geq 0$ (resp. $\nu(F) \leq 0$, $\nu(F) = 0$) for ***all*** $\mathscr{B}$***-measurable subset*** of E (i.e. $F \in \mathscr{B}$ such that $F \subseteq E$).

  In other word, E is ν***-positive***, ν***-negative***, ν***-null*** if and only if $\nu(E \cap M) > 0$, $\nu(E \cap M) < 0$, $\nu(E \cap M) = 0$ ***for any*** M. Thus if $\nu(E) = \int_X f \mathbb{1}\{E\}\, d\mu$, then it corresponds to $f \geq 0$, $f \leq 0$ and $f = 0$ for μ***-almost everywhere*** $x \in E$.

- **Lemma 1.1** *[Folland, 2013]*
*Any **measureable subset** of a positive set is positive, and the **union** of any **countable** positive set is positive.*

- **Theorem 1.2** *(**The Hahn Decomposition Theorem**)[Folland, 2013]*
*If ν is a **signed measure** on $(X, \mathscr{B})$, there exists a **positive set** P and a **negative set** N for ν such that $P \cup N = X$ and $P \cap N = \emptyset$. If P', N' is another such pair, then $P \Delta P' = N \Delta N'$ is **null** w.r.t. ν.*

- **Definition** [Folland, 2013, Resnick, 2013]
The decomposition of $X = P \cup N$ as X is a ***disjoint union*** *of a* ***positive set*** *and a* ***negative set*** is called a ***Hahn decomposition for*** ν.

- **Remark** Note that the Hahn decomposition is usually ***not unique*** as the ν-null set can be transferred between subparts P and N. To find unique decomposition, we need the following concepts:

- **Definition** [Folland, 2013]
Two *signed measures* μ, ν on $(X, \mathscr{B})$ are ***mutually singular***, or that ν is ***singular*** w.r.t. to μ, or vice versa, if and only if there exists a ***partition*** $E, F \in \mathscr{B}$ of X such that $E \cap F = \emptyset$ and $E \cup F = X$, E ***is null for*** μ and F ***is null for*** ν. Informal speaking, ***mutual singular*** means that μ and ν "***live on disjoint sets***". We describe it using perpendicular sign

$$\mu \perp \nu$$

- **Theorem 1.3** *(**The Jordan Decomposition Theorem**)[Folland, 2013]*

If ν is a signed measure on $(X, \mathscr{B})$, there exists ***unique positive measure*** *ν_+ and ν_- such that*

$$\nu = \nu_+ - \nu_- \qquad \textit{and} \qquad \nu_+ \perp \nu_-.$$

- **Definition** The two positive measures ν_+, ν_- are called the ***positive*** and ***negative variations*** of ν, and $\nu = \nu_+ - \nu_-$ is called the ***Jordan decomposition*** of ν.

  Furthermore, define the ***total variations*** of ν as the measure $|\nu|$ such that

$$|\nu| = \nu_+ + \nu_-.$$

- **Proposition 1.4** *Let ν, μ be signed measures on $(X, \mathscr{B})$ and $|\nu|$ is the total variations of ν. Then*

  1. *$E \in \mathscr{B}$ is ν-null if and only if $|\nu|(E) = 0$*
  2. *$\nu \perp \mu$* ***if and only if*** *$|\nu| \perp \mu$ if and only if $(\nu_+ \perp \mu) \wedge (\nu_- \perp \mu)$.*

- **Proposition 1.5** *If ν_1, ν_2 are signed measures that both omit $\pm\infty$, then $|\nu_1 + \nu_2| \leq |\nu_1| + |\nu_2|$*

1.2 Lebesgue Decomposition and Radon-Nikodym derivative

- **Definition** [Folland, 2013]
  Suppose ν is *a* ***signed measure*** on $(X, \mathscr{B})$ and μ is *a* ***positive measure*** on $(X, \mathscr{B})$. Then ν is said to be ***absolutely continuous w.r.t.*** μ and write

$$\nu \ll \mu$$

  if $\nu(E) = 0$ for *every $E \in \mathscr{B}$ for which $\mu(E) = 0$.*

- **Proposition 1.6** *Suppose ν is a signed measure on $(X, \mathscr{B})$, ν_+, ν_- are positive and negative variation of ν and $|\nu|$ is the total variation. Then $\nu \ll \mu$* ***if and only if*** *$|\nu| \ll \mu$* ***if and only if*** *$(\nu_+ \ll \mu) \wedge (\nu_- \ll \mu)$.*

- **Remark** ***Absolutly continuity*** is in a sense ***antithesis*** (i.e. *direct opposite*) of ***mutual singularity***. More precisely, if $\nu \perp \mu$ and $\nu \ll \mu$, then $\nu = 0$, since E, F are disjoint sets such that $E \cup F = X$, and $\mu(E) = |\nu|(F) = 0$, then $\nu \ll \mu$ implies that $|\nu|(E) = 0$. One can *extend* the notion of absolute continuity to the case where μ *is a signed measure* (namely, $\nu \ll \mu$ iff $\nu \ll |\mu|$), but we shall have no need of this more general definition.

- **Theorem 1.7** *(ϵ-δ* ***Language of Absolute Continuity of Measures****)*
  Let ν is a ***finite signed measure*** *and μ is a* ***positive*** *measure on $(X, \mathscr{B})$. Then $\nu \ll \mu$ if and only if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $|\nu(E)| < \epsilon$,* ***whenever*** *$\mu(E) < \delta$.*

- **Remark** If μ is a *measure* and f is ***extended μ-integrable***, then ***the signed measure*** ν *defined via $\nu(E) = \int_E f d\mu$ is* ***absolutely continuous*** *w.r.t.* μ; it is ***finite*** if and only if f is ***absolutely integrable***. For any complex-valued $f \in L^1(\mu)$, the preceding theorem can be applied to $\Re(f)$ and $\Im(f)$.

- **Corollary 1.8** *If $f \in L^1(X, \mu)$, for every $\epsilon > 0$, there exists a $\delta > 0$, such that $\left|\int_E f d\mu\right| < \epsilon$ whenever $\mu(E) < \delta$.*

- **Definition** For *a* ***signed measure*** ν defined via $\nu(E) = \int_E f d\mu$ for all $E \in \mathscr{B}$, we use the notation to express the relationship
$$d\nu = f\, d\mu.$$
Sometimes, by a slight abuse of language, we shall refer to "***the signed measure*** $f\, d\mu$"

- **Lemma 1.9** *[Folland, 2013]*
Suppose that ν and μ are ***finite measures*** *on $(X, \mathscr{B})$. Either $\nu \perp \mu$, or there exists $\epsilon > 0$ and $E \in \mathscr{B}$ such that $\mu(E) > 0$ and $\nu \geq \epsilon\mu$ on E, i.e. E is a* ***positive set for*** *$\nu - \epsilon\mu$.*

- **Theorem 1.10** ***(Lebesgue-Radon-Nikodym Theorem)****[Folland, 2013]*
Let ν be a ***σ-finite signed*** *measure and μ be a* ***σ-finite positive*** *measure on $(X, \mathscr{B})$. There exists* ***unique σ-finite signed measure*** *λ, ρ on $(X, \mathscr{B})$ such that*
$$\lambda \perp \mu\,, \quad \text{and} \quad \rho \ll \mu\,, \quad \text{and} \quad \nu = \lambda + \rho.$$
In particular, if $\nu \ll \mu$, then
$$d\nu = f d\mu, \qquad \text{for some } f.$$

- **Definition** The decomposition $\nu = \rho + \lambda$, where $\lambda \perp \mu$ and $\rho \ll \mu$, is called the ***Lebesgue decomposition*** *of ν with respect to μ.*

- **Remark** By Lebesgue decomposition, *a signed measure ν* can be represented as
$$d\nu = d\lambda + f d\mu$$

- **Definition** If $\nu \ll \mu$, then according to *the Lebesgue-Radon-Nikodym theorem*, $d\nu = f d\mu$ for some f, where f is called the ***Radon-Nikodym derivative*** *of ν w.r.t. μ* and is denoted as
$$f := \frac{d\nu}{d\mu} \quad \Rightarrow \quad d\nu = \frac{d\nu}{d\mu} d\mu.$$

- We stated it in terms a theorem in probability space:

  Theorem 1.11 ***(Radon-Nikodym Theorem)*** *[Resnick, 2013]*
  Let $(\Omega, \mathscr{F}, \mathcal{P})$ be the ***probability space****. Suppose ν is a* ***positive bounded measure*** *and $\nu \ll \mathcal{P}$. Then there exists an* ***$\mathscr{F}$-integrable random variable*** *X, such that*
  $$\nu(E) = \int_E X d\mathcal{P},\ \forall\, E \in \mathscr{F}.$$
  X is ***almost everywhere unique*** *(P) and is written*
  $$f = \frac{d\nu}{d\mathcal{P}}$$
  We also write $d\nu = X d\mathcal{P}$.

- **Corollary 1.12** ***(σ-Finite Measures)*** *[Resnick, 2013]*
If μ, ν are ***σ-finite measures*** *on $(\Omega, \mathscr{F})$, there exists a* ***$\mathscr{F}$-measurable*** *X such that*
$$\nu(E) = \int_E X d\mu,\ \forall\, E \in \mathscr{F},$$
if and only if
$$\nu \ll \mu.$$

- The following corollary is very important in definition of ***conditional expectation***:

  Corollary 1.13 *(**Restriction to Sub σ-Algebra**) [Resnick, 2013]*
  Suppose Q, P are both probability measure on $(\Omega, \mathscr{F})$ such that $Q \ll P$. Let $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra. Let $Q|_{\mathscr{G}}$, $P|_{\mathscr{G}}$ be the restriction of Q, P to $\mathscr{G}$. Then in $(\Omega, \mathscr{G})$,

  $$Q|_{\mathscr{G}} \ll P|_{\mathscr{G}}$$

  and

  $$\frac{d\, Q|_{\mathscr{G}}}{d\, P|_{\mathscr{G}}} \text{ is } \mathscr{G}\text{-measureable.}$$

- **Remark** (***Jordan Decomposition vs. Lebesgue Decomposition***)
  We see ***two unique decompositions***: the Jordan decomposition and the Lebesgue decomposition. We can make a comparison:

  1. Both of these two are *decompositions* of *a **signed** measure ν.*
  2. Both of these two decompositions seperate ν into two ***mutually signular*** sub-measures of ν.
  3. Both of these two decompositions are ***unique***

  On the other hand,

  1. ***The Jordan decomposition*** is to split *a signed measure ν **itself*** into ***two positive measures***, i.e. ν_+ and ν_- that are ***mutually singular*** $(\nu_+ \perp \nu_-)$.
  2. ***The Lebesgue decomposition*** is to split *a signed measure ν **with respect to a postive measure** μ*. The result is *two-fold*: 1) *two mutually singular sub-measures $\lambda \perp \rho$* 2) their relationship with μ is ***opposite***: $\lambda \perp \mu$, i.e. their support do not overlap; $\rho \ll \mu$, i.e. its support lies within support of μ.
  3. Note that λ, ρ from *the Lebesgue decomposition* is ***not*** *necessarily* ***positive***. But both ν and μ need to be ***σ-finite*** which is *not required* for the Jordan decomposition.

- **Proposition 1.14** *[Folland, 2013]*
  *Suppose ν is **σ-finite signed measure** and λ, μ are **σ-finite measure** on $(X, \mathscr{B})$ such that $\nu \ll \mu$ and $\mu \ll \lambda$.*

  1. *If $g \in L^1(X, \nu)$, then $g\left(\frac{d\nu}{d\mu}\right) \in L^1(X, \mu)$ and*

  $$\int g d\nu = \int g \frac{d\nu}{d\mu} d\mu$$

  2. *We have $\nu \ll \lambda$, and*

  $$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu}\frac{d\mu}{d\lambda}, \quad \lambda\text{-a.e.}$$

- **Corollary 1.15** *If $\mu \ll \lambda$ and $\lambda \ll \mu$, then $(d\lambda/d\mu)(d\mu/d\lambda) = 1$ a.e. (with respect to either λ or μ).*

- **Proposition 1.16** *If $\mu_1, \ldots, \mu_n$ are measures on $(X, \mathscr{B})$, then there exists a measure μ such that $\mu_i \ll \mu$ for all $i = 1, \ldots, n$, namely, $\mu = \sum_{i=1}^n \mu_i$.*

2 Conditional Probability

2.1 Definitions

- **Remark** (***Conditional probability in terms of Decision with Partial Information***)
  It is helpful to consider ***conditional probability*** in terms of *an observer in possession of **partial information***. A *probability space* $(\Omega, \mathscr{F}, \mathcal{P})$ describes the working of a ***mechanism***, *governed by chance*, which produces a result ω *distributed* according to $\mathcal{P}$; $\mathcal{P}(A)$ is for the observer the probability that the point ω produced lies in A.

  Suppose now that ω lies in G and that the observer *learns this fact and no more*. From the point of view of the *observer*, now in possession of *this partial information about* ω, *the **probability** that* ω *also lies in* A is $\mathcal{P}(A|B)$ rather than $\mathcal{P}(A)$. This is the idea lying back of the definition.

- **Remark** (***Conditional Probability with respect to Experiments***)
  Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ is a ***sub-σ-algebra*** on Ω. One can imagine an observer who knows for each G in $\mathscr{G}$ whether $w \in G$ or $w \in G^c$. Thus the σ-algebra $\mathscr{G}$ can in principle be ***identified*** with an ***experiment*** or ***observation***.

  It is natural to try and define *conditional probabilities* $\mathcal{P}[A\,|\,\mathscr{G}]$ with respect to the experiment $\mathscr{G}$. To do this, *fix* an A in $\mathscr{F}$ and define *a finite measure* ν on $\mathscr{G}$ by

  $$\nu(G) = \mathcal{P}(A \cap G), \quad G \in \mathscr{G}$$

  Then $\mathcal{P}(G) = 0$ implies that $\nu(G) = 0$, i.e. $\nu \ll \mathcal{P}$. The *Lebesgue-Radon-Nikodym Theorem* can be applied to the measures ν and $\mathcal{P}$ on the measurable space $(\Omega, \mathscr{G})$ because the first one is ***absolutely continuous*** with respect to the second. It follows that there exists a ***function*** or ***random variable*** f, $\mathscr{G}$*-measurable* and *integrable* with respect to $\mathcal{P}$, such that

  $$\nu(G) = \mathcal{P}(A \cap G) = \int_G f d\mathcal{P}$$

  for all $G \in \mathscr{G}$. This random variable f is ***the conditional probabilty of*** A ***given*** $\mathscr{G}$.

- **Definition** (***Conditional Probability***)
  Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ is a ***sub-σ-algebra*** on Ω. Given a $\mathscr{F}$***-measurable set*** $A \in \mathscr{F}$, there exists a ***random variable***, ***denoted*** *as* $\mathcal{P}[A|\mathscr{G}]$ with two properties:

  1. $\mathcal{P}[A|\mathscr{G}]$ is a $\mathscr{G}$***-measurable function*** and ***integrable*** *with respect to* $\mathcal{P}$
  2. $\mathcal{P}[A|\mathscr{G}]$ satisfies ***the functional equation***:

  $$\int_G \mathcal{P}[A|\mathscr{G}]\, d\mathcal{P} = \mathcal{P}(A \cap G), \quad \forall\, G \in \mathscr{G}.$$

  The random variable $\mathcal{P}[A|\mathscr{G}]$ is called ***the conditional probability of*** A ***given*** $\mathscr{G}$.

- **Remark** By definition, the conditional probability is a ***Radon-Nikodym derivative*** of ν w.r.t. $\mathcal{P}$.

  $$\mathcal{P}[A|\mathscr{G}] := \frac{d\nu|_{\mathscr{G}}}{d\mathcal{P}|_{\mathscr{G}}}$$

is a ***$\mathscr{G}$-measurable function***. It is ***not a measure itself*** but there exists an ***isomorphism*** $\mathcal{P}[A|\mathscr{G}] \mapsto \nu$ between a ***conditional probability*** (as random variable) given $\mathscr{G}$ and *a **probability measure** on $\mathscr{G}$.*

- **Remark** (***Observe Outcome via Functional*** $\mathcal{P}[A|\mathscr{G}]_{\omega}$)
Condition (1) in the definition above in effect requires that the ***values of*** $\mathcal{P}[A|\mathscr{G}]$ ***depend only on the sets in*** $\mathscr{G}$. An observer who knows the outcome of $\mathscr{G}$ viewed as an experiment *knows for each* G in *whether it contains* ω *or not*; *for each* x he ***knows*** this in particular for the set
$$\{\omega : \mathcal{P}[A|\mathscr{G}]_{\omega} = x\}\,,$$
and hence he knows in principle ***the functional value*** $\mathcal{P}[A|\mathscr{G}]_{\omega}$, even if he ***does not know*** ω ***itself***.

- **Remark** Note that
$$\int_{G} \mathcal{P}[A|\mathscr{G}]\, d\mathcal{P}$$
is a *measure* of $G \in \mathscr{G}$, *not a measure of* $A \in \mathscr{F}$.

- **Remark** (***σ-Algebra Generated by Partition of Sample Space***)
If $\mathscr{G}$ is the σ-algebra ***generated*** by a ***partition*** $B_1, B_2, \ldots$, then the general element of $\mathscr{G}$ is a ***disjoint union***
$$B_1 \cup \ldots \cup B_n \cup \ldots$$
finite *or* ***countable***, of certain of the B_i; To know which set B_i it is that *contains* ω is the same thing as to know which sets in $\mathscr{G}$ contain ω and which do not. This second way of looking at the matter *carries over to the general σ-algebra $\mathscr{G}$ contained in $\mathscr{F}$.*

  As always, the probability space is $(\Omega, \mathscr{F}, \mathcal{P})$. The σ-algebra $\mathscr{F}$ will not in general come from a partition as above. Then ***the conditional distribution*** $\mathcal{P}[A|\mathscr{G}]$ can be written as
$$f(\omega) := \mathcal{P}(A|B_i) = \frac{\mathcal{P}(A \cap B_i)}{\mathcal{P}(B_i)}, \quad \text{if } \omega \in B_i,\, i = 1\,, \ldots, n \ldots$$
In this case, $\mathcal{P}[A|\mathscr{G}]$ is the ***function*** whose value on B_i is *the* ***ordinary conditional probability*** $\mathcal{P}[A|B_i]$. If the observer learns which element B_i of the *partition* it is that contains ω, then his ***new probability*** for the event $\omega \in A$ is $f(\omega)$. The partition $\{B_i\}$, or equivalently the σ-algebra, $\mathscr{G}$, can be regarded as an ***experiment***, and *to **learn which** B_i **it is that contains** ω is to **learn the outcome of the experiment**.* Any G in $\mathscr{G}$ is a disjoint union $G = \bigcup_k B_{i_k}$, and
$$\begin{aligned}\mathcal{P}(A \cap G) &= \sum_{k} \mathcal{P}(A \,|\, B_{i_k})\mathcal{P}(B_{i_k}) \\ \Rightarrow \mathcal{P}[A|\mathscr{G}] &= \sum_{k} \mathcal{P}(A \,|\, B_{i_k})\mathbb{1}\,\{B_i\}\end{aligned}$$

- **Remark** (***Condition Probability Given*** $\sigma(X)$)
The σ-algebra $\sigma(X)$ generated by a random variable X consists of the sets
$$\{\omega : X(\omega) \in H\}$$

for $H \in \mathscr{B}$. The ***conditional probability of A given X*** is defined as $\mathcal{P}[A|\sigma(X)]$ and is denoted $\mathcal{P}[A|X]$. Thus

$$\mathcal{P}[A|X] := \mathcal{P}[A|\sigma(X)]$$

by definition.

- **Example** (***Discrete Case***)
  Let X be a ***discrete random variable*** with possible values $x_1, x_2, \ldots$. Then for $A \in \mathscr{F}$,

$$\begin{aligned}\mathcal{P}(A \,|\, X) &= \mathcal{P}(A \,|\, \sigma(X)) \\ &= \mathcal{P}(A \,|\, \sigma([X = x_i], i = 1, 2, \ldots)) \\ &= \sum_{i=1}^{\infty} \mathcal{P}(A \,| X = x_i) \mathbb{1}\left\{[X = x_i]\right\}.\end{aligned}$$

- **Example** (***Absolutely Continuous Case***)
  Let $\Omega = \mathbb{R}^2$ and suppose X and Y are random variables whose *joint distribution is **absolutely continuous*** with ***density*** $f(x, y)$ so that for $A \in \mathcal{B}(\mathbb{R}^2)$,

$$\mathcal{P}[(X, Y) \in A] = \iint_A f(x, y) dx dy.$$

We use $\mathscr{G} = \sigma(X)$. Let

$$I(x) := \int f(x, y) dy.$$

be the ***marginal density*** of X and define and

$$\phi(X) = \begin{cases} \frac{\int_C f(X,y)dy}{I(x)} & \text{if } I(x) > 0 \\ 0 & \text{if } I(x) = 0. \end{cases}$$

Then we claim that for $C \in \mathcal{B}(\mathbb{R})$,

$$\mathcal{P}(Y \in C|X) = \mathcal{P}(Y \in C|\sigma(X)) = \phi(X).$$

Proof: First of all, note that $\int_C f(X, y) dy$ is $\sigma(X)$-*measurable* and hence $\phi(X)$ is $\sigma(X)$-*measurable*. So it remains to show for any $\Lambda \in \sigma(X)$ that

$$\int_\Lambda \phi(X) d\mathcal{P} = \mathcal{P}([Y \in C] \cap \Lambda).$$

Since $\Lambda \in \sigma(X)$, the form of Λ is $\Lambda = [X \in A]$ for some $A \in \mathcal{B}(\mathbb{R})$. By the Transformation Theorem,

$$\begin{aligned}\int_\Lambda \phi(X) d\mathcal{P} &= \int_{X^{-1}(A)} \phi(X) d\mathcal{P} \\ &= \int_A \phi(x)\, \mathcal{P}[X \in dx]\end{aligned}$$

and because a density exists for the joint distribution of (X, Y), we get this equal to

$$\begin{aligned}
\int_A \phi(x)\, \mathcal{P}[X \in dx] &= \int_A \phi(x) \left[\int_{\mathbb{R}} f(x,y)dy\right] dx \\
&= \int_{A\cap\{x:I(x)>0\}} \phi(x) \left[\int_{\mathbb{R}} f(x,y)dy\right] dx + \int_{A\cap\{x:I(x)=0\}} \phi(x) \left[\int_{\mathbb{R}} f(x,y)dy\right] dx \\
&= \int_{A\cap\{x:I(x)>0\}} \phi(x) \left[\int_{\mathbb{R}} f(x,y)dy\right] dx + 0 \\
&= \int_{A\cap\{x:I(x)>0\}} \frac{\int_C f(x,y)dy}{I(x)} I(x)dx \\
&= \int_{A\cap\{x:I(x)>0\}} \int_C f(x,y)dydx \\
&= \int_A \int_C f(x,y)dydx = \mathcal{P}[X \in A, Y \in C] \\
&= \mathcal{P}([Y \in C] \cap \Lambda). \quad \blacksquare
\end{aligned}$$

- **Remark**

$$\mathcal{P}[X \in H|\mathscr{G}] = \mathcal{P}\left[\{\omega' : X(\omega') \in H\}\, |\mathscr{G}\right]$$

2.2 Properties

- **Proposition 2.1** ***(Conditional Probabilty from Generating π-System)****[Billingsley, 2008]*
Let $\mathscr{P}$ be a ***π-system*** *generating the σ-algebra $\mathscr{G}$, and suppose that Ω is a finite or countable* ***union*** *of sets in $\mathscr{P}$. An integrable function f is* ***a version*** *of $\mathcal{P}[A|\mathscr{G}]$ if it is* ***$\mathscr{G}$-measurable*** *and if*

$$\int_G f d\mathcal{P} = \mathcal{P}(A \cap G)$$

holds for all G in $\mathscr{P}$.

- **Remark** (***$\mathscr{G}$ as Borel σ-algebra on Subspace***)
Consider a topological space X with ***Borel σ-algebra*** $\mathscr{B}$ generated by all open sets in X, we define a ***subspace*** $S \subseteq X$ equipped with ***the subspace topology***. Then $\mathscr{G} \subset \mathscr{B}$ is ***the Borel σ-algebra on*** S. Note that a subset $G \subset S$ is *open* in S if and only if there exists some open subset $G_X \subset X$ such that

$$G = G_X \cap S.$$

Thus $\mathscr{G}$ is generated by subsets of form $(G_X \cap S)$. Thus a measure ν can be defined as the restriction of measure $\mathcal{P}$ in probability space $(X, \mathscr{B}, \mathcal{P})$ in the subspace $(S, \mathscr{G})$, so that given $A \subset X$

$$\nu(G) = \int_G \mathcal{P}[A|\mathscr{G}]d\mathcal{P} = \mathcal{P}(A \cap G) = \int_{G_X\cap S} \mathcal{P}[A|\mathscr{G}]d\mathcal{P}.$$

for all $G \in \mathscr{G}$ as subset of S.

- **Proposition 2.2** *[Billingsley, 2008]*
  With probability 1, $\mathcal{P}[\emptyset|\mathscr{G}] = 0$, $\mathcal{P}[\Omega|\mathscr{G}] = 1$*; and*
  $$0 \le \mathcal{P}[A|\mathscr{G}] \le 1$$
  for each A*. If* $A_1, A_2, \ldots$ *is a finite or countable sequence of* ***disjoint*** *sets, then*
  $$\mathcal{P}\left[\bigcup_{n=1}^{\infty} A_n \middle| \mathscr{G}\right] = \sum_{n=1}^{\infty} \mathcal{P}\left[A_n|\mathscr{G}\right].$$
  with probability 1.

2.3 Conditional Probability Distributions

- **Proposition 2.3** ***(Conditional Probability Distribution)*** *[Billingsley, 2008]*
  Let $(\Omega, \mathscr{F}, \mathcal{P})$ *be a probability space and* $\mathscr{G} \subset \mathscr{F}$ *is a* ***sub-σ-algebra*** *on* Ω*. Define* $X : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *as a random variable and* $\mathcal{B}(\mathbb{R})$ *is the* ***Borel σ-algebra*** *on* $\mathbb{R}$*. There exists a function called* ***transition function*** *or* ***transition kernel***
  $$K : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, 1]$$
  such that

  1. *For each* ω *in* Ω*,* $K(\omega, \cdot)$ *is a* ***probability measure*** *on* $\mathcal{B}(\mathbb{R})$*;*
  2. *For each* ***Borel set*** $H \in \mathcal{B}(\mathbb{R})$*,* $K(\cdot, H)$ *is the* ***conditional probability*** $\mathcal{P}[X \in H|\mathscr{G}]$*.*

  The probability measure $\mu := K(\omega, \cdot)$ *is a* ***conditional distribution of X given $\mathscr{G}$****. If* $\mathscr{G} = \sigma(Z)$*, it is a* ***conditional distribution of X given Z****.*

- **Remark** Note that ***the first argument*** of *kernel* is a ***point*** in Ω while ***the second argument*** is a ***Borel measurable set*** in $\mathcal{B}(\mathbb{R})$. Thus it make sense for
  $$K(\omega, dx) = \lim_{r \to 0} K(\omega, B(x, r)).$$

- **Remark** (***Conditional Probability Distribution $\neq$ Conditional Probability***)
  From the definition above, we see that conditional probability distribution is not the conditional probability:

  – ***A conditional probability*** $\mathcal{P}[X \in H|\mathscr{G}]$ is a ***$\mathscr{G}$-measurable function*** $f(\omega)$.
  $$\omega \mapsto \mathcal{P}[X \in H|\mathscr{G}]_{\omega}, \quad \textit{for fixed } H$$
  In other word, $\mathcal{P}[X \in H|\mathscr{G}]$ is a ***random variable*** *determined by the* ***conditioning*** *term* $\mathscr{G}$. If $\mathscr{G} = \sigma(Z)$, then $\mathcal{P}[X \in H|\sigma(Z)] = \mathcal{P}[X \in H|Z] = f(Z)$ is ***a function of conditioning random variable*** Z.

  – ***A conditional probability distribution*** is ***a measure on*** $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

  In this case, it is resulting ***probability measure*** on real line $\mathbb{R}$ when *the outcome on the* ***conditioning*** *term* ω *is* ***observed (fixed)***
  $$H \to \mathcal{P}\left[\left\{\omega' : X(\omega') \in H\right\}|\mathscr{G}\right]_{\omega}, \quad \textit{for fixed } \omega$$
  Note that the *outcome* ω in ***conditioning term*** is ***different*** from the *outcome* ω' measured in the ***preimage set*** of $X^{-1}(H)$ since they are *two* ***separated events***.

3 Conditional Expectation

3.1 Definitions

- **Definition** (***Conditional Expectation***) [Resnick, 2013]
  Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra. Suppose $X \in L^1(\Omega, \mathscr{F}, \mathcal{P})$. There exists a function $\mathbb{E}\left[X|\mathscr{G}\right]$, called the ***conditional expectation*** of X ***with respect to*** $\mathscr{G}$ such that

  1. $\mathbb{E}\left[X|\mathscr{G}\right]$ is ***$\mathscr{G}$-measureable*** and ***integrable*** *with respect to* $\mathcal{P}$.
  2. $\mathbb{E}\left[X|\mathscr{G}\right]$ satisfies ***the functional equation***:
  $$\int_G X d\mathcal{P} = \int_G \mathbb{E}\left[X|\mathscr{G}\right] d\mathcal{P}, \quad \forall\, G \in \mathscr{G}.$$

- **Remark** To *prove the* ***existence*** of such a random variable,

  1. consider first the case of ***nonnegative*** X. Define a measure ν on $\mathscr{G}$ by
  $$\nu(G) = \int_G X d\mathcal{P} = \int_\Omega X \mathbb{1}_G \, d\mathcal{P}.$$
  This measure is *finite* because X is *integrable*, and it is ***absolutely continuous*** with respect to $\mathcal{P}$. By the *Lebesgue-Radon-Nikodym Theorem*, there is a $\mathscr{G}$-measurable function f such that
  $$\nu(G) = \int_G f d\mathcal{P}.$$
  This f has properties (1) and (2).
  2. If X is *not necessarily nonnegative*, $\mathbb{E}\left[X_+|\mathscr{G}\right] - \mathbb{E}\left[X_-|\mathscr{G}\right]$ clearly has the required properties.

- **Remark** As $\mathscr{G}$ increases, condition (1) becomes ***weaker*** and condition (2) becomes ***stronger***.

- **Remark** Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space, with $\mathscr{G} \subset \mathscr{F}$ a sub-σ-algebra, define
  $$\mathcal{P}[A|\mathscr{G}] = \mathbb{E}\left[\mathbb{1}_A|\mathscr{G}\right]$$
  for all $A \in \mathscr{F}$.

- **Remark** By definition, the conditional expectation is a ***Radon-Nikodym derivative*** of $d\nu|_{\mathscr{G}} = X d\mathcal{P}|_{\mathscr{G}}$ w.r.t. $d\mathcal{P}|_{\mathscr{G}}$ within $\mathscr{G}$.
  $$\mathbb{E}\left[X|\mathscr{G}\right] := \frac{X d\mathcal{P}|_{\mathscr{G}}}{d\mathcal{P}|_{\mathscr{G}}} = X|_{\mathscr{G}}.$$
  Thus $\mathbb{E}\left[X|\mathscr{G}\right]$ *is the* ***projection*** *of X on* ***sub σ-algebra*** $\mathscr{G}$.

- **Remark** (***Conditioning on Random Variables***)
  By definition, conditioning on random variables $(X_t, t \in T)$ on $(\Omega, \mathscr{B})$ can be expressed as
  $$\mathbb{E}\left[X|X_t, t \in T\right] \equiv \mathbb{E}\left[X|\sigma(X_t, t \in T)\right],$$

where $\sigma(X_t, t \in T)$ is the σ-algebra generated by the cylinder set

$$C_n[A] \equiv \{\omega : (X_t(\omega), 1 \le t \le n) \in A\} \in \mathscr{B}, \quad A \in \mathcal{B}(\mathbb{R}^n), \forall\, n$$

- **Remark** (σ-***Algebra Generated by Partition of Sample Space***)
As above, assume that the sub σ-algebra $\mathscr{G}$ is generated by a ***partition*** $B_1, B_2, \ldots$ of Ω, then for $X \in L^1(\Omega, \mathscr{F}, \mathcal{P})$,

$$\mathbb{E}\left[X|B_i\right] = \int X d\mathcal{P}(X|B_i) = \int_{B_i} X d\mathcal{P}/\mathcal{P}(B_i)$$

where $\mathcal{P}(X|B_i)$ is *the conditional probability* defined in previous section. If $\mathcal{P}(B_i) = 0$, then $\mathbb{E}\left[X|B_i\right] = 0$. We claim that

1. $$\mathbb{E}\left[X|\mathscr{G}\right] = \sum_{i=1}^{\infty} \mathbb{E}\left[X|B_i\right] \mathbb{1}_{B_i}, \quad a.s.$$

2. For any $A \in \mathscr{F}$,
$$\mathcal{P}(A|\mathscr{G}) = \sum_{i=1}^{\infty} \mathcal{P}(A|B_i) \mathbb{1}_{B_i}, \quad a.s.$$

- **Remark** Both $P[A|\mathscr{F}]$ and $\mathbb{E}\left[X|\mathscr{F}\right]$ are *random variables* from $\Omega \to \mathbb{R}$. Formally speaking,

$$\begin{aligned} P\left[(X,Y) \in A|\sigma(X)\right]_{\omega} &\equiv P\left[(X(\omega), Y) \in A\right] \\ &= P\left\{\omega' : (X(\omega), Y(\omega')) \in A\right\} \\ &\equiv f(X(\omega)) \\ &= \nu|_{\sigma(X)}(A) \\ \mathbb{E}\left[(X,Y)|\sigma(X)\right]_{\omega} &= \lim_{\substack{m(A)\to 0 \\ \omega \in A \in \sigma(X)}} \frac{P\left\{\omega' : (X(\omega), Y(\omega')) \in A\right\}}{m(A)} \end{aligned}$$

It is the expected value of X for someone who knows for each $E \in \mathscr{F}$, whether or not $\omega \in E$, which E itself remains unknown.

3.2 Properties

- **Proposition 3.1** ***(Properties of Conditional Expectation)*** *[Resnick, 2013]*
Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra. Suppose $X, Y \in L^1(\Omega, \mathscr{F}, \mathcal{P})$ and $\alpha, \beta \in \mathbb{R}$.

1. ***(Linearity)****:* $\mathbb{E}\left[\alpha X + \beta Y|\mathscr{G}\right] = \alpha\mathbb{E}\left[X|\mathscr{G}\right] + \beta\mathbb{E}\left[Y|\mathscr{G}\right]$;
2. ***(Projection)****: If X is* ***$\mathscr{G}$-measurable****, then* $\mathbb{E}\left[X|\mathscr{G}\right] = X$ *almost surely.*
3. ***(Conditioning on Indiscrete σ-Algebra)****:*
$$\mathbb{E}\left[X|\{\emptyset, \Omega\}\right] = \mathbb{E}\left[X\right].$$

4. ***(Monotonicity)**: If $X \geq 0$, then $\mathbb{E}[X|\mathscr{G}] \geq 0$ almost surely. Similarly, if $X \geq Y$, then $\mathbb{E}[X|\mathscr{G}] \geq \mathbb{E}[Y|\mathscr{G}]$ almost surely.*

5. ***(Modulus Inequality)**:*

$$|\,\mathbb{E}[X|\mathscr{G}]\,| \leq \mathbb{E}[\,|X|\,|\mathscr{G}].$$

6. ***(Monotone Convergence Theorem)**: If $\{X_n\}_{n=1}^{\infty} \subset L^1(\Omega, \mathscr{F}, \mathcal{P})$, $0 \leq X_1 \leq X_2 \leq \dots$ is a **monotone sequence** of **non-negative** random variables and $X_n \to X$ then*

$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathscr{G}] = \mathbb{E}\left[\lim_{n\to\infty} X_n \middle| \mathscr{G}\right] = \mathbb{E}[X|\mathscr{G}].$$

7. ***(Fatou Lemma)**: If $\{X_n\}_{n=1}^{\infty} \subset L^1(\Omega, \mathscr{F}, \mathcal{P})$, and $X_n \geq 0$ for all n, then*

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n \middle| \mathscr{G}\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n|\mathscr{G}]$$

8. ***(Dominated Convergence Theorem)**: If $\{X_n\}_{n=1}^{\infty} \subset L^1(\Omega, \mathscr{F}, \mathcal{P})$ and $|X_n| \leq Z$, where $Z \in L^1(\Omega, \mathscr{F}, \mathcal{P})$ is a random variable, $X_n \to X$ almost surely, then*

$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathscr{G}] = \mathbb{E}\left[\lim_{n\to\infty} X_n \middle| \mathscr{G}\right] = \mathbb{E}[X|\mathscr{G}], \quad a.s.$$

9. ***(Product Rule)**: If Y is $\mathscr{G}$-measurable,*

$$\mathbb{E}[X\,Y|\mathscr{G}] = Y\,\mathbb{E}[X|\mathscr{G}], \quad a.s.$$

***Proof:** For any $E \in \mathscr{F}$,*

$$\begin{aligned}\int_E Y\mathbb{E}[X|\mathscr{F}]\,dP &= \int_E XY\,dP \\ &= \int_E \mathbb{E}[XY|\mathscr{F}]\,dP\end{aligned}$$

using the fact that $Y = \mathbb{1}\{A\}$ with linearity, and monotone converging theorem,

$$\begin{aligned}\int_E \mathbb{1}\{A\}\,\mathbb{E}[X|\mathscr{F}]\,dP &= \int_{E\cap A} \mathbb{E}[X|\mathscr{F}]\,dP \\ &= \int_{E\cap A} X\,dP \\ &= \int_E \mathbb{1}\{A\}\,X\,dP \quad \blacksquare\end{aligned}$$

10. ***(Smoothing)**: For $\mathscr{F}_1 \subset \mathscr{F}_0 \subset \mathscr{F}$,*

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|\mathscr{F}_0]\,|\mathscr{F}_1] &= \mathbb{E}[X|\mathscr{F}_1] \\ \mathbb{E}[\mathbb{E}[X|\mathscr{F}_1]\,|\mathscr{F}_0] &= \mathbb{E}[X|\mathscr{F}_1].\end{aligned}$$

*Note that $\mathbb{E}[X|\mathscr{F}_1]$ is **smoother** than $\mathbb{E}[X|\mathscr{F}_0]$. Moreover*

$$\mathbb{E}[X] = \mathbb{E}[X|\{\emptyset,\Omega\}] = \mathbb{E}[\mathbb{E}[X|\mathscr{F}_0]\,|\{\emptyset,\Omega\}] = \mathbb{E}[\mathbb{E}[X|\mathscr{F}_0]].$$

Proof: *Since for any $F \in \mathscr{F}_1 \subset \mathscr{F}_0 \subset \mathscr{F}$,*

$$\begin{aligned}\int_F \mathbb{E}\left[\mathbb{E}\left[X|\mathscr{F}_0\right]|\mathscr{F}_1\right]dP &= \int_F \mathbb{E}\left[X|\mathscr{F}_0\right]dP \\ &= \int_F XdP \quad (\text{ since } F \in \mathscr{F}_0 \subset \mathscr{F}) \\ &= \int_F \mathbb{E}\left[X|\mathscr{F}_1\right]dP; \quad (\text{ since } F \in \mathscr{F}_1 \subset \mathscr{F}) \quad \blacksquare\end{aligned}$$

11. *(**The Conditional Jensen's Inequality**). Let ϕ be a **convex** function, $\phi(X) \in L^1(\Omega, \mathscr{F}, \mathcal{P})$. Then almost surely*

$$\phi\left(\mathbb{E}\left[X|\mathscr{G}\right]\right) \le \mathbb{E}\left[\phi(X)|\mathscr{G}\right]$$

- **Theorem 3.2** *(**Projection Theorem** or **The Minimum Mean Squared Estimation**) [Billingsley, 2008, Resnick, 2013]*
  *Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra. $L^2(\Omega, \mathscr{G})$ is the space of the square integrable $\mathscr{G}$**-measurable** functions. If $X \in L^2(\Omega, \mathscr{F})$, then $\mathbb{E}[X|\mathscr{G}]$ is the **orthogonal projection** of X onto $L^2(\Omega, \mathscr{G}) \subset L^2(\Omega, \mathscr{F})$. That is, $\mathbb{E}[X|\mathscr{G}]$ is the **unique element** in the **subspace** $L^2(\Omega, \mathscr{G})$ that achieves*

  $$\inf_{Z \in L^2(\Omega, \mathscr{G})} \|X - Z\|_{L^2}.$$

  *In other word, $\mathbb{E}[X|\mathscr{G}]$ is **the minimum mean squared estimator (MMSE)** of X in $L^2(\Omega, \mathscr{G})$.*

  Proof: It is computed by solving the orthogonality condition for $Z \in L^2(\Omega, \mathscr{G})$:

  $$\langle Y\,,\, X - Z\rangle = 0, \quad \forall Y \in L^2(\Omega, \mathscr{G}).$$

  This says that

  $$\int Y(X - Z)d\mathcal{P} = 0, \quad \forall Y \in L^2(\Omega, \mathscr{G}).$$

  But trying a solution of $Z = \mathbb{E}[X|\mathscr{G}]$, we get

  $$\begin{aligned}\int Y(X - Z)d\mathcal{P} &= \int Y\left(X - \mathbb{E}\left[X|\mathscr{G}\right]\right)d\mathcal{P} \\ &= \mathbb{E}\left[Y\left(X - \mathbb{E}\left[X|\mathscr{G}\right]\right)\right] \\ &= \mathbb{E}\left[YX\right] - \mathbb{E}\left[Y\,\mathbb{E}\left[X|\mathscr{G}\right]\right] \\ &\quad (\text{ since } Y \text{ is } \mathscr{G}\text{-measurable}) \\ &= \mathbb{E}\left[YX\right] - \mathbb{E}\left[\mathbb{E}\left[YX|\mathscr{G}\right]\right] \\ &= \mathbb{E}\left[YX\right] - \mathbb{E}\left[YX\right] = 0. \quad \blacksquare\end{aligned}$$

- **Remark** The result above is essentially ***the projection theorem*** for Hilbert space. Note that $L^2(\Omega, \mathscr{F}, \mathcal{P})$ is a ***Hilbert space*** and $L^2(\Omega, \mathscr{G}) \subset L^2(\Omega, \mathscr{F})$ is a ***closed subspace***. Thus for every $X \in L^2(\Omega, \mathscr{F}, \mathcal{P})$, it can be ***uniquely*** written as

  $$X = \mathbb{E}\left[X|\mathscr{G}\right] + \Delta$$

  where $\Delta \perp L^2(\Omega, \mathscr{G})$.

- **Proposition 3.3** ***(Conditioning and Independence)*** *[Resnick, 2013]*
Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra. Suppose $X \in L^1(\Omega, \mathscr{F}, \mathcal{P})$.

  1. *If $X \perp\!\!\!\perp \mathscr{G}$, i.e. X is **independent** from $\mathscr{G}$, then*

$$\mathbb{E}\left[X|\mathscr{G}\right] = \mathbb{E}\left[X\right]$$

  2. *Let $\phi : \mathbb{R}^j \times \mathbb{R}^k \to \mathbb{R}$ be a **bounded Borel function**. Suppose also that $X : \Omega \to \mathbb{R}^j$, $Y : \Omega \to \mathbb{R}^k$, X is $\mathscr{G}$-measurable and Y is **independent** of $\mathscr{G}$. Define*

$$f_\phi(x) = \mathbb{E}\left[\phi(x, Y)\right].$$

  Then

$$\mathbb{E}\left[\phi(X, Y)|\mathscr{G}\right] = f_\phi(X).$$

- **Proposition 3.4** ***(Continuity in L^p Norm)*** *[Resnick, 2013]*
Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a probability space and $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra. $X \in L^p(\Omega, \mathscr{F}, \mathcal{P})$ for $1 \le p < \infty$, i.e.

$$\left\|X\right\|_p = \left(\int |X|^p \, d\mathcal{P}\right)^{1/p} < \infty.$$

Then

$$\left\|\mathbb{E}\left[X|\mathscr{G}\right]\right\|_p \le \left\|X\right\|_p,$$

*and conditional expectation $\mathbb{E}\left[X|\mathscr{G}\right]$ as functional of X is **continuous** in L^p **norm topology**, i.e.*

$$X_n \overset{L^p}{\to} X \quad \textit{implies} \quad \mathbb{E}\left[X_n|\mathscr{G}\right] \overset{L^p}{\to} \mathbb{E}\left[X|\mathscr{G}\right]$$

- **Proposition 3.5** ***(Conditional Distributions and Expectations)*** *[Billingsley, 2008]*
*Let $K(\omega, \cdot)$ be a **conditional distribution** with respect to $\mathscr{G}$ of a random variable X, in the sense of Proposition 2.3. If $\varphi : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function for which $\varphi(X)$ is **integrable**, then*

$$\int_{\mathbb{R}} \varphi(x) K(\omega, dx) = \mathbb{E}\left[\varphi(X)|\mathscr{G}\right]_\omega, \quad a.s.$$

- **Remark** (***Conditional Expectation as Expectation w.r.t. Conditional Probability***)
It is a consequence of the proof above that $\int_{\mathbb{R}} \varphi(x) K(\omega, dx)$ is ***$\mathscr{G}$-measurable*** and ***finite*** *with probability* 1. If X is itself integrable, it follows by the $\varphi(x) = x$ that

$$\begin{aligned}
\mathbb{E}\left[X|\mathscr{G}\right]_\omega &= \int_{\mathbb{R}} x K(\omega, dx), \quad a.s. \\
&= \int_{\mathbb{R}} x \, P(dx|\mathscr{G})_\omega.
\end{aligned}$$

References

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

Sidney I Resnick. *A probability path*. Springer Science & Business Media, 2013.