Radon-Nikodym Theorem and Conditional Expectation

February 13, 2002

Conditional expectation reflects the change in unconditional probabilities due to some auxiliary information. The latter is represented by a sub-σ-algebra $\mathcal{G}$ of the basic σ-algebra of an underlying probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Note that, the *conditional expectation* of random variable X, given the σ-algebra $\mathcal{G}$, denoted by $\mathsf{E}(X|\mathcal{G})$, is itself a ($\mathcal{G}$-measurable) random variable.

1 Some preliminary functional analysis

Let X be a vector space. A **norm** on X is a function $\|\cdot\| : X \to [0, \infty)$ such that

1. $\|x\| = 0$ if and only if $x = 0$.
2. $\|x + y\| \leq \|x\| + \|y\|$, for all $x, y \in X$.
3. $\|\lambda x\| = |\lambda| \cdot \|x\|$, for all $x \in X$.

The space $(X, \|\cdot\|)$ is said to be a **normed vector space**. The norm naturally defines a metric on space X.

Definition: A normed vector space $(X, \|\cdot\|)$ is said to be a **Banach space** if it is complete with respect to the norm-metric.

Remark: A sequence $\{x_n\} \subseteq X$ is said to be a *cauchy sequence* if for any $\epsilon > 0$, there exists $N \in \mathsf{N}$ such that $\|x_m - x_n\| \leq \epsilon$ for all $m, n \geq N$. A metric space is *complete* if any Cauchy sequence converges.

Lemma: A normed vector space $(X, \|\cdot\|)$ is complete if and only if for every sequence $\{x_n\}_{n\in\mathsf{N}} \subseteq X$ with property

$$\sum_{n=1}^{\infty} \|x_n\| < \infty,$$

the sequence $S_n \doteq \sum_{j=1}^{n} x_j$ converges.

Proof. "⇒". Suppose X is complete, and $\sum_{n=1}^{\infty} \|x_n\| < \infty$. Then with $S_n \doteq \sum_{j=1}^{n} x_j$, it is easy to see that $\{S_n\}$ is a Caushy sequence since

$$\|S_n - S_m\| = \|\sum_{j=m}^{n} x_j\| \leq \sum_{j=m}^{n} \|x_j\| \to 0, \quad \text{as } m, n \to \infty.$$

Hence $\{S_n\}$ converges since X is complete.

"$\Leftarrow$". Let $\{x_n\} \subseteq X$ be a Cauchy sequence. We can find $n_1 < n_2 < \cdots$ such that

$$\|x_n - x_m\| \leq \frac{1}{2^j}, \quad \forall \ m, n \geq n_j.$$

Let $y_1 = x_{n_1}$, $y_j = x_{n_j} - x_{n_{j-1}}$ $(j \geq 2)$. It follows that

$$\sum_{j=1}^{\infty} \|y_j\| \leq \|y_1\| + \sum_{j=2}^{\infty} \|x_{n_j} - x_{n_{j-1}}\| \leq \|y_1\| + \sum_{j=1}^{\infty} \frac{1}{2^j} < \infty.$$

By assumption, the sequence $\left\{x_{n_k} = \sum_{j=k}^{n} y_j\right\}$ converges, say to $x \in X$. Since $\{x_n\}$ is a Cauchy sequence, we have $\lim_n x_n = x$ (for every $\epsilon > 0$, choose $k \in \mathsf{N}$ such that $\|x_{n_j} - x_{\|} \leq \epsilon$ for all $j \geq k$, and $l \geq k$ such that $2^{-l} \leq \epsilon$. Then $\|x_n - x\| \leq \|x_n - x_{n_l}\| + \|x_{n_l} - x\| \leq 2^{-l} + \epsilon \leq 2\epsilon$). □

Proposition: For any $1 \leq p < \infty$, the space $\mathsf{L}^p(\mu)$ is a Banach space with the L^p-norm.

Proof. We should use the lemma above. Consider a sequence $\{f_n\} \subseteq \mathsf{L}^p$ with $\sum_n \|f_n\|_p = S$ being finite. Let $G_n \doteq \sum_{j=1}^{n} |f_j|$. It follows that $G_n \uparrow \sum_{j=1}^{\infty} |f_j| := G$. It follows from MCT that

$$\|G\|_p = \lim_n \|G_n\|_p \leq \lim_n \sum_{j=1}^{n} \|f_j\|_p \leq S < \infty.$$

Hence $G \in \mathsf{L}^p$; in particular, G is finite almost everywhere, or $\sum_{j=1}^{\infty} |f_j|$ converge almost everywhere. Define $F = \sum_{j=1}^{\infty} f_j$, which exists almost everywhere. Of course $F \in \mathsf{L}^p$ since $|F| \leq G$. It remaines to show that

$$\|F - \sum_{j=1}^{n} f_j\|_p \to 0.$$

However,

$$\left| F - \sum_{j=1}^{n} f_j \right|^p =\leq \left(|F| + \sum_{j=1}^{\infty} |f_j| \right)^p \leq (2G)^p \in \mathsf{L}^1.$$

It follows from DCT that

$$\lim_n \|F - \sum_{j=1}^{n} f_j\|_p^p = \|F - \lim_n \sum_{j=1}^{n} f_j\|_p^p = 0,$$

this completes the proof. □

The most import Banach space is the so-called Hilbert space. Suppose X is a vector space. An **inner product** is a function $(x, y) \mapsto \langle x, y \rangle$ from $X \times X$ into R, such that

1. $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$ for all $a, b \in \mathsf{R}$, $x, y \in X$.
2. $\langle x, y \rangle = \langle y, x \rangle$, for all $x, y \in X$.
3. $\langle x, x \rangle \geq 0$, for all $x \in X$, with equality if and only if $x = 0$.

Cauchy-Bunyakowsky-Schwarz Inequality. If $\langle \cdot, \cdot \rangle$ is an inner product on vector space X, then

$$\langle x, y \rangle \leq \langle x, x \rangle \langle y, y \rangle$$

for all $x, y \in X$.

Proof. For any $t \geq 0$ and $x, y \in X$, we have

$$0 \leq \langle x - ty, x - ty \rangle = \langle x, x \rangle - 2t\langle x, y \rangle + t^2 \langle y, y \rangle.$$

If $\langle y, y \rangle = 0$, then $y = 0$ and the inequality automaticaly holds. Otherwise, the above quardratic form (with respect to t) is always non-negative if and only if its discriminant

$$4\langle x, y \rangle^2 - 4\langle x, x \rangle \langle y, y \rangle \leq 0.$$

This completes the proof. □

Exercise: The mapping $X \to [0, \infty)$ with $x \mapsto \sqrt{\langle x, x \rangle} := \|x\|$ defines a norm on vector space X.

Exercise **(Paralleogram Law):** For any $x, y \in X$, we have

$$\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right)$$

Definition: If $(X, \|\cdot\|)$ is complete, we say $(X, \langle \cdot, \cdot \rangle)$ is a **Hilbert space**.

Example: The space L^2 with inner product

$$\langle f, g \rangle \doteq \int f g \, d\mu, \qquad \forall \ f, g \in \mathsf{L}^2$$

is a Hilbert space.

Exercise: Suppose X is a Hilbert space, and $T : X \to \mathsf{R}$ is a linear functional. The following statements are equivalent.

1. T is continuous;
2. T is continuous at some point;
3. There exists a constant c such that $|T(x)| \leq c\|x\|$ for all $x \in X$.

Proof. It is clear that $1 \Leftrightarrow 2$ and $3 \Rightarrow 1$. It remains to show that $1 \Rightarrow 3$. The continuity implies that there exists a $\delta > 0$ such that $|T(x)| < 1$ whenever $\|x\| < \delta$. Now for an arbitrary $x \in X$ and $\epsilon > 0$, we have

$$\left\| \frac{\delta x}{\|x\| + \epsilon} \right\| < \delta \quad \Rightarrow \quad \frac{\delta}{\|x\| + \epsilon} |T(x)| < 1;$$

that is

$$|T(x)| < \frac{\|x\| + \epsilon}{\delta}, \quad \forall \ x \in X, \ \epsilon > 0.$$

Lettint $\epsilon \to 0$, we complete the proof. □

Definition: Two elements $x, y \in X$ are **orthogonal** if $\langle x, y\rangle = 0$. For any $Y \subseteq X$, its **orthogonal complement** $Y^\perp$ is defined as

$$Y^\perp \doteq \{x \in X;\ \langle x, y\rangle = 0,\ \ \forall\ y \in Y\}$$

Exercise: For an arbitrary $Y \subseteq X$, $Y^\perp$ is a closed sub-vector space.

Projection theorem: Suppose Y is a closed sub-vector space of X. Then $X = Y \oplus Y^\perp$; i.e., for every $x \in X$, there exists $y \in Y$, $z \in Y^\perp$ such that $x = y + z$, and this "direct sum" decomposition is unique.

Proof. Fix $x \in X$, define

$$\delta \doteq \inf_{y \in Y} \|x - y\|.$$

We claim that the infimum is achieved, say at y^*, and $x - y^* \in Y^\perp$, which implies $x = y^* + (x - y^*)$. Indeed, consider a minimizing sequence $\{y_n\} \subseteq Y$. The Paralleogram Law implies that

$$\begin{aligned}
\|y_n - y_m\|^2 &= 2(\|y_n - x\|^2 + \|y_m - x\|^2) - \|y_n + y_m - 2x\|^2 \\
&= 2(\|y_n - x\|^2 + \|y_m - x\|^2) - 4\left\|\frac{y_n + y_m}{2} - x\right\|^2 \\
&\leq 2(\|y_n - x\|^2 + \|y_m - x\|^2 - 2\delta^2) \to 0
\end{aligned}$$

as $m, n \to \infty$. It follows that $\{y_n\}$ is a Cauchy sequence, whence it converges, say to y^*. It follows that $y^* \in Y$ (thanks to the closedness of Y) achieves the infimum with $\delta = \|x - y^*\|$. It remains to show that $x - y^* \in Y^\perp$: for any $z \in Y$, and $t \in \mathsf{R}$,

$$y^* + tz \in Y \quad \Rightarrow \quad \delta^2 \leq \|x - y^* - tz\|^2 = \|x - y^*\|^2 + t^2\|z\|^2 - 2t\langle x - y^*, z\rangle := f(t).$$

Note that $f(t)$ attaines minimum at $t = 0$, we have $f'(0) = 0$ or $\langle x - y^*, z\rangle = 0$. This yields that $x - y^* \in Y^\perp$. The uniqueness is trivial. □

Riesz Representation Theorem: If $T : X \to \mathsf{R}$ is a continuous linear functional, then there exists a unique $x_0 \in X$ such that $T(x) = \langle x, x_0\rangle$ for all $x \in X$.

Proof. If $T(x) \equiv 0$, then we can choose $x_0 = 0$. Otherwise, let $Y \doteq T^{-1}(0)$. It follows that Y is a closed sub-vector space (why?). Since $Y \neq X$, $Y^\perp \neq \{0\}$. Therefore, there exists a $z \in Y^\perp$ such that $T(z) = 1$. Now for any $x \in X$, $x - T(x)z \in Y$. It follows that

$$\langle x - T(x)z, z\rangle = 0 \quad \Rightarrow \quad T(x) = \frac{1}{\|z\|^2}\langle x, z\rangle := \langle x, x_0\rangle, \ \text{ where } \ x_0 = \frac{z}{\|z\|^2}.$$

The uniqueness is trivial since if $\langle x, x_0\rangle \equiv \langle x, x_1\rangle$, it is easy to see that $\|x_0 - x_1\| = 0$ by taking $x = x_0 - x_1$. □

2 Relations between measures

Let $(\Omega, \mathcal{F})$ be a measurable space, and μ, ν two measures on it. We say that ν **is absolutely continuous with respective to** μ (write $\nu \ll \mu$) if $A \in \mathcal{F}$, $\mu(A) = 0$ implies that $\nu(A) = 0$. We say μ and ν are **equivalent** (write $\mu \sim \nu$) if $\nu \ll \mu$ and $\mu \ll \nu$. We say μ and ν are **singular** if there exists a set $A \in \mathcal{F}$ such that $\mu(A) = \nu(A^c) = 0$; in which case we write $\mu \perp \nu$.

Example: Suppose $f \geq 0$ is a measurable function. The mapping $\nu : E \mapsto \int_E f \, d\mu$ defines a measure ν. It is not difficult to see that $\nu \ll \mu$. Indeed, this example is not so special. Later we should see that the reverse is also true under some very mild conditions – the Radon-Nikodým theorem.

Example: Suppose $\Omega = [0, 1]$, C is the Cantor set and $f : [0, 1] \to [0, 1]$ is the Cantor function. Note C is a compact set with $\lambda(C) = 0$; here λ denotes the Lebesgue-measure. The Cantor function f is continuous, non-decreasing and flat on set $\Omega \backslash C$; $f(0) = 0$, $f(1) = 1$. Let μ_f denote the measure induced by f on $[0, 1]$. It is not difficult to see that $\mu_f(\Omega \backslash C) = 0 = \lambda(C)$. In other words, μ_f and λ are singular.

Lebesgue Decomposition Theorem: Let $(\Omega, \mathcal{F})$ be a measurable space and μ, ν two σ-finite measures on it. Then there exist measure ν_{ac} and ν_{s} such that

$$\nu = \nu_{\text{ac}} + \nu_{\text{s}}; \qquad \nu_{\text{ac}} \ll \mu, \quad \nu_{\text{s}} \perp \mu.$$

This decomposition is unique.

Example: For example, let $\lambda\big|_E$ denote the Lebesgue measure restricted on set E; that is $\lambda\big|_E(A) \doteq \lambda(A \cap E)$ for all $A \in \mathcal{F}$. Suppose $\mu = \lambda\big|_{[0,2]}$ and $\nu = \lambda\big|_{[1,3]}$. Then

$$\nu_{\text{ac}} = \lambda\big|_{[1,2]}, \qquad \nu_{\text{s}} = \lambda\big|_{[2,3]}$$

Radon-Nikodým Theorem: Let $(\Omega, \mathcal{F})$ be a measurable space and μ, ν two σ-finite measures on it, with $\nu \ll \mu$. Then there exists a unique (up to a.e. equivalence) measurable function $h : \Omega \to [0, \infty)$ such that

$$\nu(A) = \int_A h \, d\mu, \qquad \forall \ A \in \mathcal{F}.$$

The function h is called **Radon-Nikodým derivative** of ν with respect to μ, and we write

$$h = \left.\frac{d\nu}{d\mu}\right|_{\mathcal{F}} := \frac{d\nu}{d\mu}.$$

Exercise: If $\nu \ll \mu$, then

$$\int f \, d\nu = \int f \frac{d\nu}{d\mu} \, d\mu$$

whenever either integral is well-defined.

*Exercise (***Chain Rule***):* If $\xi \ll \nu \ll \mu$, then

$$\frac{d\xi}{d\mu} = \frac{d\xi}{d\nu} \cdot \frac{d\nu}{d\mu}, \qquad \text{a.e. } (\mu).$$

Proof. We will first prove the case where both μ and ν are finite. Let $X \doteq \mathsf{L}^2(\Omega, \mathcal{F}, \mu+\nu)$, which is a Hilbert space if we define the inner product by $\langle f, g\rangle \doteq \int fg\, d(\mu+\nu)$. Consider the mapping

$$T:\ X \to \mathsf{R}, \quad f \mapsto \int f\, d\nu.$$

It is easy to see that T is a continuous mapping, since

$$\big|T(f)\big| \le \int |f|\, d(\mu+\nu) \le \sqrt{\nu(\Omega)+\mu(\Omega)}\sqrt{\|f\|_2}$$

by Hölder inequality. It follows from Riesz representation that there exists a unique $g \in X$ such that

$$(*) \qquad T(f) = \int f\, d\nu = \int fg\, d(\mu+\nu) \quad \Rightarrow \quad \int f\, d\mu = \int f(1-g)\, d(\mu+\nu), \quad \forall\ f \in X.$$

We claim that $0 \le g \le 1$, $(\mu+\nu)$-almost everywhere. Indeed, let $f = 1_{\{g<0\}}$, we have

$$0 \le \nu\{g<0\} = \int_{\{g<0\}} g\, d(\mu+\nu) \quad \Rightarrow \quad (\mu+\nu)\{g<0\} = 0.$$

Similarly, let $f = 1_{\{g>1\}}$, we have

$$0 \le \mu\{g>1\} = \int_{\{g>1\}} (1-g)\, d(\mu+\nu) \quad \Rightarrow \quad (\mu+\nu)\{g>1\} = 0.$$

It is a direct consequence of MCT that the (*) holds for all non-negative measurable function f.

Existence of the decomposition theorem: Let $A \doteq \{\omega;\ g(\omega) = 1\}$. It follows from (*) that $\mu(A) = 0$. Define

$$\nu_{\mathrm{s}}(E) \doteq \nu(E \cap A), \qquad \nu_{\mathrm{ac}}(E) \doteq \nu(E \cap A^c), \quad \forall\ E \in \mathcal{F}.$$

Clearly ν_{s} and ν_{ac} are two measures with $\nu = \nu_{\mathrm{ac}} + \nu_{\mathrm{s}}$. Furthermore, $\nu_{\mathrm{s}} \perp \mu$ since $\nu_{\mathrm{s}}(A^c) = \mu(A) = 0$. It remains to show that $\nu_{\mathrm{ac}} \ll \mu$: for all $E \in \mathcal{F}$ such that $\mu(E) = 0$, we have, by letting $f = 1_E$,

$$0 = \mu(E) = \int_E (1-g)\, d(\mu+\nu) \quad \Rightarrow \quad (\mu+\nu)(E \cap A^c) = 0 \quad \Rightarrow \quad \nu_{\mathrm{ac}}(E) = \nu(E \cap A^c) = 0.$$

Uniqueness of the decomposition theorem: Suppose $\nu = \rho + \sigma$ is another decomposition with $\rho \ll \mu$ and $\sigma \perp \mu$. We first show that $\nu_{\mathrm{s}} \le \mu$. Indeed, since $\mu(A) = 0$, we have $\rho(E \cap A) = 0$ for all $E \in \mathcal{F}$. Hence,

$$\nu_{\mathrm{s}} = \nu(E \cap A) = \rho(E \cap A) + \sigma(E \cap A) = \sigma(E \cap A) \le \sigma(E), \quad \forall\ E \in \mathcal{F}.$$

It follows that

$$\sigma - \nu_{\mathrm{s}} = \nu_{\mathrm{ac}} - \rho$$

is a measure which is both absolutely continous and singular with respect to μ, which implies that $\sigma - \nu_{\mathrm{s}} = \nu_{\mathrm{ac}} - \rho = 0$.

Existence of the R-N theorem: Assumem that $\nu \ll \mu$. Define

$$h \doteq \left\{ \begin{array}{ccc} \frac{g}{1-g} & ; & \text{on } A^c = \{g < 1\} \\ 0 & ; & \text{on } A = \{g = 1\} \end{array} \right\}.$$

Since $\mu(A) = 0$, we have

$$\begin{aligned} \int_E h\, d\mu &= \int_{E\cap A^c} h\, d\mu = \int h 1_{E\cap A^c}(1-g)\, d(\mu+\nu) = \int g 1_{E\cap A^c}\, d(\mu+\nu) = \int 1_{E\cap A^c}\, d\nu \\ &= \nu(E\cap A^c) = \nu_{\rm ac}(E) = \nu(E). \end{aligned}$$

Uniqueness of the R-N theorem: This is trivial.

Extension to the σ-finiteness measures: It is not difficult to find a sequence of disjoint sub-spaces $\{\Omega_n\} \subseteq \mathcal{F}$ such that $(\mu+\nu)(\Omega_n) < \infty$ and $\cup_n \Omega_n = \Omega$. For any measure ρ, let ρ_n be the measure restricted on Ω_n, or $\rho_n(E) = \rho(E \cap \Omega_n)$ for all $E \in \mathcal{F}$. It follows that $\rho = \sum_n \rho_n$, and $\rho(E) = 0$ if and only if $\rho_n(E) = 0$ for all $n \in \mathsf{N}$. It is not difficult to establish that

$$\nu \ll \mu \Leftrightarrow \nu_n \ll \mu_n, \quad \forall\, n; \quad \nu \perp \mu \Leftrightarrow \nu_n \perp \mu_n, \quad \forall\, n.$$

Therefore, we can find two measures on Ω_n, $\rho^{(n)} \ll \mu_n$ and $\sigma^{(n)} \perp \mu_n$, such that $\nu_n = \rho^{(n)} + \sigma^{(n)}$. It is not difficult to see that $\nu_{\rm ac}(E) = \sum_n \rho^{(n)}(E \cap \Omega_n)$ and $\nu_{\rm s}(E) = \sum_n \sigma^{(n)}(E \cap \Omega_n)$ are the decomposition. The uniqueness follows readily, since another decomposition will have to coincide with $\rho^{(n)}$ and $\sigma^{(n)}$ on each Ω.

As for the R-N theorem, we can find $h_n : \Omega_n \to [0, \infty)$ on each Ω_n. Define $h(\omega) = h_n(\omega)$, $\forall \omega \in \Omega_n$. It follows that h is non-negative and measurable, such that

$$\int_E h\, d\mu = \sum_n \int_{E\cap\Omega_n} h_n\, d\mu = \sum_n \nu(E \cap \Omega_n) = \nu(E).$$

The uniqueness is also trivial. □

Exercise: The **relative entropy** of a probability measure ν with respect to another probability measure μ is defined as

$$H(\nu\|\mu) \doteq \left\{ \begin{array}{ccc} \int \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu}\, d\mu & ; & \text{if } \nu \ll \mu \\ \infty & ; & \text{otherwise} \end{array} \right\}.$$

The **total variation distance** (on $\mathcal{F}$) between two probability measures μ, ν are defined as

$$\|\mu - \nu\| \doteq \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|.$$

Show that

$$\|\mu - \nu\|^2 \le \frac{1}{2} H(\nu\|\mu).$$

3 Conditional Expectation

The general condition expectation is itself a random variable. Consider a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, and a random variable X defined on it. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra (intuitively, $\mathcal{G}$ is the additional knowledge you acquire).

Definition: Suppose $\mathsf{E}X$ is well defined. We say Y is the **conditional expectation of X given $\mathcal{G}$** if the following two conditions hold:

1. $Y : \Omega \to \bar{\mathsf{R}}$ is $\mathcal{G}$-measurable.
2. $\int_E Y \, d\mathsf{P} = \int_E X \, d\mathsf{P}$, for all $E \in \mathcal{G}$.

We should denote the conditional expectation Y by $\mathsf{E}(X \,|\, \mathcal{G})$. In particular, when $X = 1_A$ for some $A \in \mathcal{F}$, we sometimes write $\mathsf{P}(A \,|\, \mathcal{G}) = \mathsf{E}(X \,|\, \mathcal{G})$.

Remark: There can be many versions of $\mathsf{E}(X \,|\, \mathcal{G})$, which differ on P-null sets.

Before discussing the general existence and uniquesness (up to a.e. equivalence), and properties of conditional expectation, we would like to know the connection of conditional expection to the more conventional conditional probability.

Example: Given $A, B \in \mathcal{F}$, the conventional conditional probability is given by

$$\mathsf{P}(A \,|\, B) \doteq \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)};$$

here we assume $\mathsf{P}(B) > 0$. This definition can be understood as the probability of A when event B occurs.

Now define a sub-σ-algebra $\mathcal{G} \doteq \{\emptyset, B, B^c, \Omega\} \subseteq \mathcal{F}$. We want to compute the conditional expectation $Y = \mathsf{P}(A \,|\, \mathcal{G})$. By definition, Y is $\mathcal{G}$-measurable, hence we can write $Y = a1_B + b1_{B^c}$ for some constant $a, b \in \bar{\mathsf{R}}$. However, for any $E \in \mathcal{G} = \{\emptyset, B, B^c, \Omega\}$, we have

$$\int_E Y \, d\mathsf{P} = \int_E 1_A \, d\mathsf{P} \quad \Rightarrow \quad a\mathsf{P}(E \cap B) + b\mathsf{P}(E \cap B^c) = \mathsf{P}(A \cap E).$$

Let $E = B$ and $E = B^c$, we have

$$a = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}, \qquad b = \frac{\mathsf{P}(A \cap B^c)}{\mathsf{P}(B^c)}.$$

Moreover, when a, b are chosen as above, it is not difficult to see that

$$a\mathsf{P}(E \cap B) + b\mathsf{P}(E \cap B^c) = \mathsf{P}(A \cap E), \qquad \forall\ E \in \mathcal{G}.$$

It follows that

$$\mathsf{P}(A \,|\, \mathcal{G})(\omega) = (\text{the conventional}) \left\{ \begin{array}{lcl} \mathsf{P}(A \,|\, B) & ; & \text{if} \quad \omega \in B; \\ \mathsf{P}(A \,|\, B^c) & ; & \text{if} \quad \omega \in B^c; \end{array} \right\}$$

Therefore, the definition of $\mathsf{P}(A \,|\, \mathcal{G})$ coincides with the conventional conditional probability well. $\square$

Example: Suppose X, Y are two independent, integrable random variables on space $(\Omega, \mathcal{F}, \mathsf{P})$. The **$\sigma$-algebra generated by random variable** Z is defined as

$$\sigma(Z) \doteq \left\{ Z^{-1}(B); \quad B \in \mathcal{B}(\mathsf{R}) \right\}.$$

Let $\mathcal{G} \doteq \sigma(Y)$. What will be $\mathsf{E}(X+Y \,|\, \mathcal{G})$? Intuitively, knowing Y does not provide any additional information of X due to the independence. Hence, a candidate for this conditional expectation is $\mathsf{E}(X) + Y$. This claim is easily verified. Indeed, $\mathsf{E}(X) + Y$ is obviously $\mathcal{G}$-measurable. Secondly, for any $E = Y^{-1}(B) \in \mathcal{G}$, we have

$$\int_E (X+Y)\, d\mathsf{P} = \mathsf{E}(X 1_E + Y 1_E) = \mathsf{E}(X) \cdot \mathsf{P}(E) + \mathsf{E}(Y 1_E);$$

here the last equality follows from independence. Also,

$$\int_E \left[E(X) + Y \right] \, d\mathsf{P} = \mathsf{E}(X) \cdot \mathsf{P}(E) + \mathsf{E}(Y 1_E).$$

Therefore, the definition of conditional expectation fits the intuition. □

Now we return to the existence and uniqueness of the conditional expectation.

Theorem: Consider a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, a random variable X on it and a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. If $\mathsf{E}X$ is well-defined, then there exists a $\mathcal{G}$-measurable function $\mathsf{E}(X \,|\, \mathcal{G}) : \Omega \to \bar{\mathsf{R}}$, unique to P-null sets, such that

$$\int_E \mathsf{E}(X \,|\, \mathcal{G})\, d\mathsf{P} = \int_E X \, d\mathsf{P}, \quad \forall\ E \in \mathcal{G}.$$

We call $\mathsf{E}(X \,|\, \mathcal{G})$ the **conditional expectation of X given $\mathcal{G}$**. In particular,

$$\mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{G})\big) = \mathsf{E}X.$$

Remark: When X is integrable; i.e. $\mathsf{E}|X| < \infty$, the conditional expectation $\mathsf{E}(X \,|\, \mathcal{G})$ is finite almost surely. This is implied in the proof of the theorem.

Proof. We give here a proof for the case where $X \in \mathsf{L}^1$. Suppose $X \in \mathsf{L}^1$, then $X^{\pm} \in \mathsf{L}^1$. The measures defined by

$$\nu^{\pm}(E) \doteq \int_E X^{\pm} \, d\mathsf{P}, \quad \forall\ E \in \mathcal{G},$$

are two finite measures absolutely continuous to P. By Radon-Nikodým theorem, there exist $\mathcal{G}$-measurable functions $h^{\pm} : \Omega \to [0, \infty)$ such that

$$\nu^{\pm}(E) \doteq \int_E h^{\pm} \, d\mathsf{P}, \quad \forall\ E \in \mathcal{G}.$$

Define $\mathsf{E}(X \,|\, \mathcal{G}) = h^+ - h^-$. The uniqueness is trivial. □

Exercise: Complete the proof for the case where $\mathsf{E}X$ is well-defined.

Proof. It suffice to consider the case $\mathsf{E}X = \infty$, or equivalently, $\mathsf{E}X^+ = \infty$ and $\mathsf{E}X^- < \infty$. Similarly, we can define two measures $\nu^{\pm}$ on $\mathcal{G}$, and ν^- can be taken care of as in the above proof. It remains to show that there exists a $\mathcal{G}$-measurable function $h^+ : \Omega \to [0, \infty]$ such that

$$\nu^+(E) = \int_E X^+ \, d\mathsf{P} = \int_E h^+ \, d\mathsf{P}, \qquad \forall \ E \in \mathcal{G}.$$

Note in this case, we still have $\nu^+ \ll \mathsf{P}$, but the σ-finiteness of ν^+ is not guaranteed. We define

$$\mathcal{D} \doteq \big\{ D \in \mathcal{G}; \quad \nu^+ \ \text{ is } \sigma\text{-finite on } \mathcal{G} \cap D. \big\}$$

Define $\alpha \doteq \sup_{D \in \mathcal{D}} \mathsf{P}(D)$. Select a sequence of $\{D_n\} \in \mathcal{D}$ such that $\mathsf{P}(D_n) \uparrow \alpha$, and define $F \doteq \cup_n D_n$. We have $\mathsf{P}(F) = \alpha$ and $F \in \mathcal{D}$; in other words, $\mu \doteq \nu^+\big|_{\mathcal{G} \cap F}$ is σ-finite. We claim that for any $A \in \mathcal{G} \cap F^c$, either $\mathsf{P}(A) = \nu^+(A) = 0$, or $\mathsf{P}(A) > 0$ and $\nu^+(A) = \infty$. Indeed, if $\mathsf{P}(A) > 0$ and $\nu^+(A) < \infty$, we have $A \cup F \in \mathcal{D}$, and $\mathsf{P}(A \cup F) = \mathsf{P}(A) + \alpha > \alpha$, a contradiction. Now define

$$h^+ \doteq \left\{ \begin{array}{ccc} \frac{d\mu}{d\mathsf{P}'} & ; & \text{on } F \\ \infty & ; & \text{on } F^c \end{array} \right\}; \qquad \text{here } \mathsf{P}' = \mathsf{P}\big|_{\mathcal{G} \cap F}.$$

It remains to show that

$$\nu^+(E) = \int_{E \cap F} h^+ \, d\mathsf{P} + \int_{E \cap F^c} h^+ \, d\mathsf{P} = \int_E h \, d\mathsf{P}, \qquad \forall \ \ E \in \mathcal{G},$$

which is trivial. □

Below is a collection of exercises (we assume all the conditional expectations are well-defined in these exercises).

Exercise: Suppose $X, Y \in \mathsf{L}^1$. If Y is a $\mathcal{G}$-measurable random variable such that

$$\mathsf{E}X = \mathsf{E}Y, \quad \int_E X \, d\mathsf{P} = \int_E Y \, d\mathsf{P}, \quad \forall \ E \in \mathcal{D},$$

for some π-class $\mathcal{D}$. Then

$$Y = \mathsf{E}(X \,|\, \mathcal{G}), \qquad \text{where } \ \mathcal{G} = \sigma(\mathcal{D}).$$

Exercise: Let $\{A_n\}_{n \in \mathsf{N}} \subseteq \mathcal{F}$ be a partition of Ω; that is, $\{A_n\}_{n \in \mathsf{N}}$ is a disjoint sequence with $\cup A_n = \Omega$. Let $\mathcal{G} \doteq \sigma(A_n; \ n \in \mathsf{N})$ be the sub-σ-algebra generated by $\{A_n\}$. Suppose X is a random variable with $\mathsf{E}X$ well-defined. Show that

$$\mathsf{E}(X \,|\, \mathcal{G}) = \frac{\int_{A_n} X \, d\mathsf{P}}{\mathsf{P}(A_n)}, \qquad \text{on } A_n.$$

Note when $\mathsf{P}(A_n) = 0$, the above fraction is defined as any real number.

Exercise: For arbitrary constants $\alpha, \beta \in \mathsf{R}$, we have

$$\mathsf{E}(\alpha X + \beta Y \,|\, \mathcal{G}) = \alpha \mathsf{E}(X \,|\mathcal{G}) + \beta \mathsf{E}(Y \,|\mathcal{G}).$$

Exercise: If X is $\mathcal{G}$-measurable, then $\mathsf{E}(X \,|\, \mathcal{G}) = X$.

Exercise: Suppose $X \geq Y$ almost surely. Show that $\mathsf{E}(X \,|\, \mathcal{G}) \geq \mathsf{E}(Y \,|\mathcal{G})$ almost surely. In particular, if $X \geq 0$ almost surely, we have $\mathsf{E}(X \,|\, \mathcal{G}) \geq 0$ almost surely.

Exercise: Show that $|\mathsf{E}(X \,|\, \mathcal{G})| \leq \mathsf{E}(|X| \,\big|\, \mathcal{G})$ almost surely. (Hint: let $Y = \pm X$, and use the above exercise).

Exercise: (**Conditional Monotone Convergence Theorem**) If $0 \leq X_n \uparrow X$, then $\mathsf{E}(X_n \,|\, \mathcal{G}) \uparrow \mathsf{E}(X \,|\, \mathcal{G})$ almost surely.

Proof. Let $Y_n = \mathsf{E}(X_n \,|\, \mathcal{G})$. It follows that $\{Y_n\}$ is an increasing sequence, and hence $Y \doteq \lim_n Y_n$ exists and is $\mathcal{G}$-measurable. It follows from MCT and the definition, that for any $E \in \mathcal{G}$,

$$\int_E Y \, d\mathsf{P} = \lim_n \int_E Y_n \, d\mathsf{P} = \lim_n \int_E X_n \, d\mathsf{P} = \int_E X \, d\mathsf{P} = \int_E \mathsf{E}(X \,|\, \mathcal{G}) \, d\mathsf{P}.$$

This completes the proof. □

Exercise: (**Conditional Fatou Lemma**) If $0 \leq X_n$, then $\mathsf{E}\left(\liminf_n X_n \,|\, \mathcal{G}\right) \leq \liminf_n \mathsf{E}(X_n \,|\, \mathcal{G})$ almost surely.

Proof. Let $Y_n \doteq \inf_{m \geq n} X_m$, we have $Y_n \uparrow \liminf_n X$. It follows from CMCT that

$$\mathsf{E}\left(\liminf_n X_n \,|\, \mathcal{G}\right) = \lim_n \mathsf{E}(Y_n \,|\, \mathcal{G}) = \liminf_n \mathsf{E}(Y_n \,|\, \mathcal{G}) \leq \liminf_n \mathsf{E}(X_n \,|\, \mathcal{G})$$

Exercise: (**Conditional Dominated Convergence Theorem**) If $|X_n| \leq Y \in \mathsf{L}^1$ and $X_n \to X$ almost surely, then $\mathsf{E}(X_n \,|\, \mathcal{G}) \to \mathsf{E}(X \,|\, \mathcal{G})$ and $\mathsf{E}(|X_n - X| \,\big|\, \mathcal{G}) \to 0$ almost surely. (Hint: Mimic the proof of DCT.)

3.1 A special case

Suppose X, Y are two random variables. The **σ-algebra generate by** X is

$$\sigma(X) \doteq \left\{X^{-1}(B); \quad B \in \mathcal{B}(\mathsf{R})\right\}.$$

We define

$$\mathsf{E}(Y \,|\, X) \doteq \mathsf{E}\big(Y \,|\, \sigma(X)\big), \quad \text{if } \mathsf{E}Y \text{ is well-defined.}$$

More generally, if $\{X_n\}$ is a sequence of random variables,

$$\mathsf{E}(Y \,|\, X_1, X_2, \cdots) \doteq \mathsf{E}\big(Y \,|\, \sigma(X_1, X_2, \cdots)\big), \quad \text{if } \mathsf{E}Y \text{ is well-defined}$$

Lemma: Any $\sigma(X)$-measurable random variable Z can be written as $Z = \phi(X)$ for some Borel-measurable $\phi : \mathsf{R} \to \mathsf{R}$.

The proof of the lemma is left as an exercise. An immediate consequence of the lemma is

Theorem: There exists a Borel-measurable funciton ϕ such that $\mathsf{E}(Y \,|\, X) = \phi(X)$ almost surely.

Remark on notation: The Borel function ϕ is sometimes denoted by

$$\phi(x) = \mathsf{E}(Y \,|\, X = x);$$

Exercise: Suppose X and Y are independent random variables and $f : \mathsf{R}^2 \to \mathsf{R}$ is a Borel-measurable function. If $\mathsf{E}f(X,Y)$ is well-defined, then

$$\mathsf{E}\left[f(X,Y) \,\middle|\, X\right] = \phi(X)$$

where $\phi(x) = \mathsf{E}f(x,Y)$ for all $x \in \mathsf{R}$ if the expectation $\mathsf{E}f(x,Y)$ is well-defined, and $\phi(x) = 0$ otherwise.

Exercise: Suppose (X,Y) has a joint density $f(x,y)$. Its marginal densities are

$$f_X(x) = \int_{\mathsf{R}} f(x,y)\,dy, \qquad f_Y(y) = \int_{\mathsf{R}} f(x,y)\,dx.$$

Define the *conditional density*

$$f_{Y|X}(y|x) \doteq \left\{ \begin{array}{ccc} \frac{f(x,y)}{f_X(x)} & ; \quad \text{if} & f_X(x) > 0 \\ 0 & ; \quad \text{if} & f_X(x) = 0 \end{array} \right\}.$$

Show that for Borel-measurable function ϕ where $h(Y) \in \mathsf{L}^1$,

$$\mathsf{E}\big(h(Y) \,|\, X\big) = \phi(X)$$

where

$$\phi(x) \doteq \int_{\mathsf{R}} h(y) f_{Y|X}(y|x)\,dy.$$

4 The interpretation of conditional expectation as a projection

We first introduce the following result.

Theorem: Suppose random variable Y is $\mathcal{G}$-measurable, and $\mathsf{E}X$, $\mathsf{E}(XY)$ are both well-defined. We have

$$\mathsf{E}(XY \,|\, \mathcal{G}) = Y \cdot \mathsf{E}(X \,|\, \mathcal{G}).$$

In particular, $\mathsf{E}(XY) = \mathsf{E}\big\{Y \cdot \mathsf{E}(X \,|\, \mathcal{G})\big\}$.

Proof. Without loss of generality we assume $X \geq 0$. It suffices to show that

$$\int_E XY\,d\mathsf{P} = \int_E Y \cdot \mathsf{E}(X \,|\, \mathcal{G})\,d\mathsf{P}, \quad \forall\ E \in \mathcal{G}.$$

If $Y = 1_A$ for some $A \in \mathcal{G}$, we have

$$\int_E XY\,d\mathsf{P} = \int_{A\cap E} X\,d\mathsf{P} = \int_{A\cap E} \mathsf{E}(X \,|\, \mathcal{G})\,d\mathsf{P} = \int_E Y \cdot \mathsf{E}(X \,|\, \mathcal{G})\,d\mathsf{P}, \quad \forall\ E \in \mathcal{G}.$$

Hence the equality holds for all non-negative simple $\mathcal{G}$-measurable random variable Y. It follows from approximation and MCT that the equality holds for all non-negative $\mathcal{G}$-measurable random variable Y. In general, $Y = Y^+ - Y^-$. We complete the proof. □

Suppose $X \in \mathsf{L}^2(\Omega, \mathcal{F}, \mathsf{P}) := \mathsf{L}^2(\mathcal{F})$. Since $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-algebra, $\mathsf{L}^2(\mathcal{G}) \subseteq \mathsf{L}^2(\mathcal{F})$. Consider the following optimization problem:

$$\inf_{Y \in \mathsf{L}^2(\mathcal{G})} \mathsf{E}(X-Y)^2 = \inf_{Y \in \mathsf{L}^2(\mathcal{G})} \|X - Y\|^2_{\mathsf{L}^2}.$$

The answer to this problem is that the infimum is achieved at $Y^* = \mathsf{E}(X \,|\, \mathcal{G})$. Indeed, note that

$$\langle X - Y^*, Z \rangle = \langle X - \mathsf{E}(X \,|\, \mathcal{G}), Z \rangle = 0, \qquad \forall\ Z \in \mathsf{L}^2(\mathcal{G}),$$

thanks to the preceding theorem. It follows that for all $Y \in \mathsf{L}^2(\mathcal{G})$,

$$\begin{aligned} \mathsf{E}(X-Y)^2 &= \|X - Y^* + Y^* - Y\|^2 = \|X - Y^*\|^2 + \|Y^* - Y\|^2 + 2\langle X - Y^*, Y^* - Y \rangle \\ &= \|X - Y^*\|^2 + \|Y^* - Y\|^2 \geq \|X - Y^*\|^2 = \mathsf{E}(X - Y^*)^2. \end{aligned}$$

We can write

$$X = \mathsf{E}(X \,|\, \mathcal{G}) + [X - \mathsf{E}(X \,|\, \mathcal{G})] := Y^* + Z,$$

then Y^* is the "projection of X on $\mathcal{G}$", and $Z \perp \mathsf{L}^2(\mathcal{G})$ is the "orthogonal complement".

5 Other properties of conditional expectation

Consider a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$.

Proposition: (*Tower property*) Suppose $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$. We have

$$\mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{G}_2) \,\big|\, \mathcal{G}_1\big) = \mathsf{E}(X \,|\, \mathcal{G}_1) = \mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{G}_1) \,\big|\, \mathcal{G}_2\big).$$

Proof. The second equality is trivial, since $\mathsf{E}(X \,|\, \mathcal{G}_1)$ is $\mathcal{G}_2$-measurable. As for the first equality, note that $\mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{G}_2) \,\big|\, \mathcal{G}_1\big)$ is a $\mathcal{G}_1$-measurable with

$$\int_E \mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{G}_2) \,\big|\, \mathcal{G}_1\big)\, d\mathsf{P} = \int_E \mathsf{E}(X \,|\, \mathcal{G}_2)\, d\mathsf{P} == \int_E X\, d\mathsf{P} = \int_E \mathsf{E}(X \,|\, \mathcal{G}_1)\, d\mathsf{P}, \qquad \forall\ E \in \mathcal{G}_1.$$

The equality follows. □

Proposition: If $\mathcal{G} = \{\emptyset, \Omega\}$ or X is independent of $\mathcal{G}$, then

$$\mathsf{E}(X \,|\, \mathcal{G}) = \mathsf{E}X.$$

Proof. This is left as an exercise. □

Proposition: Suppose $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ are two sub-σ-algebra and $X \in \mathsf{L}^1$. If $\mathcal{G}$ is independent of X and $\mathcal{H}$, then

$$\mathsf{E}(X \,|\, \mathcal{G} \vee \mathcal{H}) = \mathsf{E}(X \,|\, \mathcal{H}).$$

Here $\mathcal{G} \vee \mathcal{H}$ is the smallest σ-algebra containing both $\mathcal{G}$ and $\mathcal{H}$; i.e.

$$\mathcal{G} \vee \mathcal{H} = \sigma(\mathcal{G}, \mathcal{H}) \doteq \sigma(\{A \cap B;\ A \in \mathcal{G},\ B \in \mathcal{H}\}) := \sigma(\mathcal{D}).$$

Remark: We say to σ-algebra $\mathcal{F}$ and $\mathcal{G}$ are independent, if A and B are independent whenever $A \in \mathcal{F}$, $B \in \mathcal{G}$. We say X is independent of a $\mathcal{G}$ if $\sigma(X)$ and $\mathcal{G}$ are independent. In particular, X and Y are independent if and only if $\sigma(X)$ and $\sigma(Y)$ are independent.

Proof. Without loss of generality, we assume $X \geq 0$. Note $\mathcal{D}$ is a π-class. Since X is integrable, it suffices to show that

$$\int_{A\cap B} X \, d\mathsf{P} = \int_{A\cap B} \mathsf{E}(X \,|\, \mathcal{H}) \, d\mathsf{P}, \quad \forall\ A \in \mathcal{G},\ B \in \mathcal{H}.$$

Indeed, if this equality holds, we have

$$\mathsf{E}(X \,|\, \mathcal{G} \vee \mathcal{H}) = \mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{H}) \,|\, \mathcal{G} \vee \mathcal{H}\big) = \mathsf{E}(X \,|\, \mathcal{H}),$$

thanks to the fact that $\mathsf{E}(X) = \mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{H})\big)$, and the exercise in page 10.

However, since $\mathsf{E}(X \,|\, \mathcal{H})$ is $\mathcal{H}$-measurable, it is independent of $\mathcal{G}$. Therefore,

$$\begin{aligned}
\int_{A\cap B} \mathsf{E}(X \,|\, \mathcal{H}) \, d\mathsf{P} &= \mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{H}) 1_A 1_B\big) = \mathsf{E}\big(\mathsf{E}(X \,|\, \mathcal{H}) 1_B\big) \cdot \mathsf{E}(1_A) \\
&= \mathsf{E}\big(\mathsf{E}(1_B X \,|\, \mathcal{H})\big) \cdot \mathsf{E}(1_A) = \mathsf{E}(X 1_B) \cdot \mathsf{E}(1_A) = \mathsf{E}(X 1_B \cdot 1_A) \\
&= \int_{A\cap B} X \, d\mathsf{P}.
\end{aligned}$$

This completes the proof. □

Proposition (conditional Hölder inequality): Suppose $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$. We have

$$\mathsf{E}\big(|XY| \,\big|\, \mathcal{G}\big) \leq \{\mathsf{E}(|X|^p \,|\, \mathcal{G})\}^{\frac{1}{p}} \cdot \{\mathsf{E}(|Y|^q \,|\, \mathcal{G})\}^{\frac{1}{q}}$$

for any random variables X, Y.

Proof. Without loss of generality we can assume that $X \geq 0$, $Y \geq 0$. We can also assume X, Y are bounded from above (otherwise, let $X_n = X \wedge n$, $Y_n = Y \vee n$ and use CMCT). We can further assume that X, Y are bounded from below by some $\epsilon > 0$ (otherwise, we can assume $X_n = X \vee \frac{1}{n}$, $Y_n = Y \vee \frac{1}{n}$ and use CDCT). Note $\{\mathsf{E}(X^p \,|\, \mathcal{G})\}^{\frac{1}{p}}$ and $\{\mathsf{E}(Y^q \,|\, \mathcal{G})\}^{\frac{1}{q}}$ are both positive and finite. It follows that

$$\frac{X}{\{\mathsf{E}(X^p \,|\, \mathcal{G})\}^{\frac{1}{p}}} \cdot \frac{Y}{\{\mathsf{E}(Y^q \,|\, \mathcal{G})\}^{\frac{1}{q}}} \leq \frac{1}{p} \left(\frac{X}{\{\mathsf{E}(X^p \,|\, \mathcal{G})\}^{\frac{1}{p}}} \right)^p + \frac{1}{q} \left(\frac{Y}{\{\mathsf{E}(Y^q \,|\, \mathcal{G})\}^{\frac{1}{q}}} \right)^q;$$

see the proof of Hölder inequality. Taking expectation conditional on $\mathcal{G}$ for both sides, we have

$$\frac{\mathsf{E}(XY \,|\, \mathcal{G})}{\{\mathsf{E}(X^p \,|\, \mathcal{G})\}^{\frac{1}{p}} \cdot \{\mathsf{E}(Y^q \,|\, \mathcal{G})\}^{\frac{1}{q}}} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

This completes the proof. □

Proposition (Conditional Jensen inequality): Suppose ϕ is a convex function, and $X, \phi(X) \in \mathsf{L}^1$. Then we have

$$\phi\big(\mathsf{E}(X \,|\, \mathcal{G})\big) \leq \mathsf{E}\big(\phi(X) \,|\, \mathcal{G}\big)$$

Proof. The convex function ϕ can be written as

$$\phi(x) = \sup_n (a_n + b_n x), \qquad \forall\ x \in \mathsf{R}$$

for a suitably chosen sequence of constants $\{(a_n, b_n)\}_{n\in\mathsf{N}}$ (exercise). Therefore,

$$\mathsf{E}\big(\phi(X)\,|\,\mathcal{G}\big) \geq \mathsf{E}\big(a_n + b_n X\,|\,\mathcal{G}\big) = a_n + b_n \mathsf{E}(X\,|\,\mathcal{G}), \quad \text{a.s., } \forall\ n.$$

In other words, the above inequality holds on a set $\Omega\backslash B_n$ with $\mathsf{P}(B_n) = 0$. Taking supremum over n, it is easy to see that

$$\mathsf{E}\big(\phi(X)\,|\,\mathcal{G}\big) \geq \phi\big(\mathsf{E}(X\,|\,\mathcal{G})\big),$$

holds on $\Omega\backslash B$ where $B = \cup_n B_n$ has probability zero; that is, the inequality hold almost surely. □

Exercise: Show that the convex function ϕ can be written as $\phi(x) = \sup_n(a_n + b_n x)$, $\forall\ x \in \mathsf{R}$ for a suitable choice of $\{(a_n, b_n)\}_{n\in\mathsf{N}}$.

Proof. As we pointed before, the following inequality holds:

$$f(x) \geq f(y) + (x - y)D^+ f(y), \qquad \forall\ x, y \in \mathsf{R};$$

here $D^+ f$ is the right-derivative of f. It follows easily that

$$f(x) = \sup_{y\in\mathsf{R}} \big[f(y) + (x - y)D^+ f(y)\big] := \sup_{y\in\mathsf{R}} [A_y + B_y x]\,; \quad \forall\ x \in \mathsf{R}$$

Indeed, $A_y + B_y x$ is called a *line of support*. The point of this exercise is that f can be expressed as the supremum of a countable collection of line of support. Let Q be the set of rational numbers. We claim

$$f(x) = \sup_{y\in\mathsf{Q}} [A_y + B_y x]\,, \quad \forall\ x \in \mathsf{R}.$$

The direction "$\geq$" is obvious. For $x \in \mathsf{R}$, let $\{y_n\} \subseteq \{\mathsf{Q}\}$ with $y_n \to x$. It follows that

$$A_{y_n} + B_{y_n} x = f(y_n) + (x - y_n)D^+ f(y_n) \quad \to \quad f(x)$$

since f is continuous and $\{D^+ f(y_n)\}$ is clearly bounded. We conclude the "$\leq$" direction. □

Here is a collection of exercises.

Exercise: Let $\{X_1, X_2, \cdots\}$ be a sequence of iid random variables with $\mathsf{E}|X_1| < \infty$. Define $\mathcal{G}_n \doteq \sigma(S_n, S_{n+1}, \cdots)$, where $S_n \doteq \sum_{j=1}^n X_j$. Show that

$$\mathsf{E}\left[X_1 \,\middle|\, \mathcal{G}_n\right] = \frac{S_n}{n}, \quad \text{a.s., } \forall\ n \geq 1.$$

Exercise: Give an example to show that $\mathsf{E}(Y\,|\,X) = \mathsf{E}Y$ almost surely does *not* necessarily imply that X and Y are independent.

6 Regular conditional distribution

Regular conditional probability directly address the issue of conditional distribution and thus provides a very nice way of describing conditional expectation.

Definition: Let $X : \Omega \to \mathsf{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. A **regular conditional distribution of X given $\mathcal{G}$** is a function $\mu : \Omega \times \mathcal{B}(\mathsf{R}) \to [0,1]$ with notation $\mathsf{P}^\omega \doteq \mu(\omega, \cdot) : \mathcal{B}(\mathsf{R}) \to [0,1]$, such that

1. For each $\omega \in \Omega$, P^ω defines a probability measure on $(\mathsf{R}, \mathcal{B}(\mathsf{R}))$.
2. For each $A \in \mathcal{B}(\mathsf{R})$, the mapping $\omega \mapsto \mathsf{P}^\omega(A)$ is $\mathcal{G}$-measurable.
3. For each $A \in \mathcal{B}(\mathsf{R})$, $\mathsf{P}(X \in A \,|\, \mathcal{G})(\omega) = \mathsf{E}(1_A \,|\, \mathcal{G})(\omega) = \mathsf{P}^\omega(A)$ for almost every $\omega \in \Omega$.

Remark: When $X : \Omega \to \mathsf{R}^n$ is a random vector or, more generally, a random variable taking value is a metric space $(S, \mathcal{B}(S))$, one can similarly define the regular conditional distribution.

Theorem: Regular conditional distribution always exists. This result holds for the general case when $X : \Omega \to (S, \mathcal{B}(S))$ is a general random variable, if S is a complete, seperable metric space (i.e. **Polish space**).

One advantage of the regular conditional distribution is that the conditional expectation can be expressed as the ordinary expectation relative the conditional distribution.

Theorem: Suppose X is a random variable, and let $\mu : \Omega \times \mathcal{B}(\mathsf{R}) \to [0,1]$ denote the regular conditional distribution of X given $\mathcal{G}$, with notation $\mathsf{P}^\omega \doteq \mu(\omega, \cdot) : \mathcal{B}(\mathsf{R}) \to [0,1]$. Let $h : \mathsf{R} \to \mathsf{R}$ be a Borel-measurable function such that $\mathsf{E}h(X)$ is well-defined. We have

$$\mathsf{E}\big(h(X) \,|\, \mathcal{G}\big)(\omega) = \int_{\mathsf{R}} h(x)\, d\mathsf{P}^\omega(x), \qquad \text{a.s.}$$

The proof of this theorem is left as an exercise (Hint: assume $h \geq 0$; show that it holds for simple functions; then use approximation...)