Appendix C: Conditional Expectation

We wish to extend the "elementary probability" definition of conditional probability and expectation to the following settings:

- $P(X_{T+k} \in \cdot | X_0, \ldots, X_T)$ (i.e. conditioning on a random number of rv's)
- $E[X|Z_0, Z_1, \ldots]$ (i.e. conditioning on a countably infinite number of rv's)
- $E[X|Z(u) : 0 \leq u \leq t]$ (i.e. conditioning on a continuum of rv's).

The "elementary probability" definition of conditional probability and expectation start from the joint pmf or joint pdf describing both the conditioned rv and the conditioning rv's. This approach clearly fails when there are infinitely many conditioning rv's (since the joint pmf and joint pdf typically fail to exist in the infinite dimensional setting).

Note that conditional probability is a special case of a conditional expectation, so we focus on conditional expectations.

C.1 The Prediction Problem

Suppose that we know the joint distribution of (X, Z), where X is real-valued. Given an observation of Z, how should we then predict X?

To be more specific, we say that W is a *predictor* of X based on Z if it can be represented as $W = h(Z)$, where $h(\cdot)$ is deterministic. We now wish to determine the optimal predictor. Of course, the notion of optimality depends on how we measure the prediction error. A natural measure of predictor error is $E^{1/p}[|X-W|^p]$. So, our goal is to compute the optimal predictor $\hat{X}$ that minimizes $E^{1/p}[|X - W|^p]$ over all predictors W.

C.2 L^p Spaces

For $p \geq 1$, let

$$L^p = \{W : W \text{ is a rv for which } E|W|^p < \infty\}.$$

Note that L^p is a vector space, in the sense that if $W_1, W_2 \in L^p$, then Minkowski's inequality implies that

$$aW_1 + W_2 \in L^p$$

for any (deterministic) constant $a \in \mathbb{R}$. Furthermore, if we set

$$\|W\|_p = E^{1/p}[|W|^p],$$

$\|\cdot\|_p$ is a *norm* on L^p, so that it has the following properties:

- $\|W_1 + W_2\|_p \leq \|W_1\|_p + \|W_2\|_p$ for $W_1, W_2 \in L^p$
- $\|aW\|_p = |a| \cdot \|W\|_p$, for $a \in \mathbb{R}$ and $W \in L^p$

- $\|W\|_p = 0$ iff $W = 0$.

Definition C.2.1 A sequence $(W_n : n \geq 0)$ is *Cauchy* in L^p if for each $\epsilon > 0$, there exists $N = N(\epsilon)$ such that if $m, n \geq N$, then

$$\|W_n - W_m\|_p < \epsilon$$

Definition C.2.2 A sequence $(W_n : n \geq 0)$ is said to *converge* to W_∞ in L^p if $W_\infty \in L^p$ and

$$\|W_n - W_\infty\|_p \to 0$$

as $n \to \infty$.

Remark C.2.1 We write

$$W_n \xrightarrow{L^p} W_\infty$$

if $(W_n : n \geq 0)$ converges to W_∞ in L^p. Convergence in L^p is often called "convergence in p'th mean". For the special case in which $p = 2$, such convergence is called "convergence in mean square".

The normed vector space L^p is *complete* in $\|\cdot\|_p$. By this, we mean that for every Cauchy sequence in L^p, there exists $W_\infty \in L^p$ such that the sequence converges to W_∞ in L^p; this is the Riesz-Fisher theorem.

Let

$$\mathcal{L} = \{W \in L^p : W = h(Z) \text{ for some deterministic } h(\cdot)\}.$$

Clearly, $\mathcal{L}$ is a linear subspace of L^p. It turns out that $\mathcal{L}$ is *closed* in the sense that if $W_n \in \mathcal{L}$ for $n \geq 0$ and

$$W_n \xrightarrow{L^p} W_\infty$$

as $n \to \infty$, then $W_\infty \in \mathcal{L}$. With the above terminology in hand, we can rephrase the prediction problem as follows:

Given $X \in L^p$, find the rv $\hat{X} \in \mathcal{L}$ which minimizes $\|X - W\|_p$ for all $W \in \mathcal{L}$.

Remark C.2.2 A complete normed vector space (like L^p) is called a *Banach space*.

C.3 The Hilbert Space L^2

When $p = 2$, the prediction problem simplifies (in the same way as "least squares" provides simpler computational formulas when minimizing a p-norm with $p \neq 2$). One explanation for the simplification is that one can introduce a "geometry" on L^2 (so that, for example, angles between rv's make sense).

For $W_1, W_2 \in L^2$, put $\langle W_1, W_2 \rangle = E[W_1 W_2]$.

The functional $\langle \cdot \rangle$ is an *inner product*, by which we mean that the following properties hold:

- $\langle a_1 W_1 + a_2 W_2, W_3 \rangle = a_1 \langle W_1, W_3 \rangle + a_2 \langle W_2, W_3 \rangle$ for $a_1, a_2 \in \mathbb{R}$, $W_1, W_2 \in L^2$
- $\langle W_1, W_2 \rangle = \langle W_2, W_1 \rangle$

- $\langle W, W\rangle$ is a norm.

Remark C.3.1 The Cauchy-Schwarz inequality implies that $|\langle W_1, W_2\rangle| \leq \|W_1\|_2 \cdot \|W_2\|_2$ for $W_1, W_2 \in L^2$.

Remark C.3.2 We can define the angle θ between $W_1, W_2 \in L^2$ via the relation

$$\cos\theta = \frac{\langle W_1, W_2\rangle}{\|W_1\|_2\|W_2\|_2}.$$

In particular, W_1 and W_2 are *orthogonal* rv's if $\langle W_1, W_2\rangle$ $(= EW_1W_2) = 0$.

Remark C.3.3 A Banach space that is equipped with an inner product is called a *Hilbert space*. So, L^2 is a Hilbert space.

C.4 The Hilbert Space Projection Theorem

For $p = 2$, we can apply the Hilbert space projection theorem to characterize the best predictor $\hat{X}$.

Theorem C.4.1 Let $\mathcal{L}$ be a closed subspace of L^2. For any $X \in L^2$, there exists a unique minimizer $\hat{X} \in \mathcal{L}$ to the problem

$$\min_{w\in\mathcal{L}} \|X - W\|_2.$$

The minimizer $\hat{X}$ is characterized by the orthogonality relationship

$$\langle X - \hat{X}, W\rangle = 0$$

for $W \in \mathcal{L}$.

Hence, the best "mean square predictor" $\hat{X}$ of X is the rv $\hat{X} \in \mathcal{L}$ for which

$$E[XW] = E[\hat{X}W] \tag{C.4.1}$$

for all $W \in \mathcal{L}$.

Definition C.4.1 If $X \in L^2$, we define the *conditional expectation of X given Z*, written as $E(X|Z)$, to be the best mean square predictor of X. In other words, $E(X|Z)$ is the rv in $\mathcal{L}$ for which

$$E[E(X|Z)h(Z)] = E[Xh(Z)] \tag{C.4.2}$$

for all deterministic $h(\cdot)$ for which $h(Z) \in L^2$.

This defines the conditional expectation for $X \in L^2$. Actually, we can generalize it further, as follows:

Step 1: If X is non-negative, put $X_n = X \wedge n$ (where $a \wedge b \triangleq \min\{a, b\}$). Clearly, $X_n \in L^2$, so $E(X_n|Z)$ is well-defined. It is straightforward to establish that $(E(X_n|Z) : n \geq 0)$ is a non-decreasing sequence of rv's, so

$$\lim_{n\to\infty} E(X_n|Z)$$

exists a.s. Define $E(X|Z)$ to be that limit

$$\text{i.e.} \quad E(X|Z) = \lim_{n\to\infty} E(X_n|Z).$$

Step 2: Suppose X is of mixed sign, but $E|X| < \infty$. Write $X = X_+ - X_-$, where $X_+ = X \vee 0$ (where $a \vee b \triangleq \max\{a, b\}$) and $X_- = (-X) \vee 0$. According to Step 1, $E(X_+|Z)$ and $E(X_-|Z)$ are well-defined. Since $E|X| = E(X_+ + X_-) < \infty$, evidently $EX_+ = EE(X_+|Z) < \infty$, so $E(X_+|Z) < \infty$ a.s. Similarly, $E(X_-|Z) < \infty$ a.s. Define $E(X|Z)$ as follows:

$$E(X|Z) = E(X_+|Z) - E(X_-|Z). \tag{C.4.3}$$

Note that "$\infty - \infty$" does not occur here, so (C.4.3) is a well-defined rv.

Hence, this extends the definition of conditional expectation to both non-negative and integrable rv's.

Remark C.4.1 For X non-negative or $X \in L^1$, $E(X|Z)$ can alternatively be characterized through the following generalization of (C.4.1):

$$EXh(Z) = EE(X|Z)h(Z) \tag{C.4.4}$$

for all bounded (deterministic) $h(\cdot)$ (or, when $X \geq 0$ via the requirement that

$$EXh(Z) = EE(X|Z)h(Z) \tag{C.4.5}$$

for all non-negative (deterministic) $h(\cdot)$).

C.5 The Connection to the Elementary Definition

Consider the special case in which (X, Z) has a joint pdf $(f(x, z) : x \in \mathbb{R}, z \in \mathbb{R}^d)$. The elementary definition of $E(X|Z = z)$ is:

$$E(X|Z = z) = \frac{\int_{-\infty}^{\infty} x f(x, z) dx}{\int_{-\infty}^{\infty} f(x, z) dx};$$

note that $E(X|Z = z)$ assigns a value (or number) to each $z \in \mathbb{R}^d$.

On the other hand, our general definition of $E(X|Z)$ defines the conditional expectation as a rv. This raises the question: How is the general definition a legitimate generalization of the elementary definition?

The connection is that in this setting,

$$E(X|Z) = h(Z) \text{ a.s.},$$

where $h(\cdot)$ is the deterministic function given by,

$$h(z) = E(X|Z = z).$$

So, $E(X|Z)$ does indeed generalize the elementary definition; an identical relationship holds, when (X, Z) has a joint pmf.

C.6 Basic Properties of Conditional Expectation

By working directly from either of the characterizations (C.4.4) or (C.4.5), it is readily shown that if $X_1, X_2 \in L^1$, then

- $E(aX_1 + X_2|Z) = aE(X_1|Z) + E(X_2|Z)$ (for $a \in \mathbb{R}$)
- $E(h(Z)X|Z) = h(Z)E(X|Z)$ for all (deterministic) h for which $E(h(Z)X|Z)$ is well-defined
- $E(X|Z_1) = E(E(X|Z_1, Z_2)|Z_1)$
- $EX = EE(X|Z)$

C.7 Other Equivalent Approaches to Defining $E(X|Z)$

Rather than following the approach described here (of first defining X for $X \in L^2$ and then extending it to $X \in L^1$), one can directly establish the existence of a rv $E(X|Z)$ satisfying (C.4.4) by appealing to the Radon-Nikodym theorem; this was the idea used by Kolmogorov to define $E(X|Z)$ in the early 20th century, and was one of his critical contributions towards putting probability on a rigorous foundation.