Chapter 4

Dependent Random Variables

Conditioning

One of the central concepts in probability theory is the notion of conditional probabiliy and conditional expectation. Suppose that we have a probaility space $(\Omega, \mathcal{F}, P)$ consisting of a space Ω, a σ-field $\mathcal{F}$ of subsets of Ω and a probability measure on the σ-field $\mathcal{F}$. If we have a set $A \in \mathcal{F}$ of positive measure then conditioning with respect to A means we restrict ourselves to the set A. Ω gets replaced by A. The σ-field $\mathcal{F}$ by the σ-field $\mathcal{F}_A$ of subsets of A that are in $\mathcal{F}$. For $B \subset A$ we define

$$P_A(B) = \frac{P(B)}{P(A)}$$

We could achieve the same thing by defining for arbitrary $B \in \mathcal{F}$

$$P_A(B) = \frac{P(A \cap B)}{P(A)}$$

in which case $P_B(\cdot)$ is a measure defined on $\mathcal{F}$ as well but one that is concentrated on A and assigning 0 probability to A^c. The definition of conditional probability is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Similarly the definition of conditional expectation of an integrable function $f(\omega)$ given a set $A \in \mathcal{F}$ of positive probability is defined to be

$$E\{f|A\} = \frac{\int_A f(\omega)dP}{P(A)}$$

In particular if we take $f = \chi_B$ for some $B \in \mathcal{F}$ we recover the definition of conditional probability. In general if we know $P(B|A)$ and $P(A)$ we can recover $P(A \cap B) = P(A)P(B|A)$ but we cannot recover $P(B)$. But if we know $P(B|A)$ as well as $P(B|A^c)$ along with $P(A)$ and $P(A^c) = 1 - P(A)$ then

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(A)P(B|A) + P(A^c)P(B|A^c)$$

More generally if $\mathcal{P}$ is a partition of Ω into a finite or even a countable number of disjoint measurable sets $A_1, \cdots, A_j, \cdots$

$$P(B) = \sum_j P(A_j)P(B|A_j)$$

If ξ is a random variable taking distinct values $\{a_j\}$ on $\{A_j\}$ then

$$P(B|\xi = a_j) = P(B|A_j)$$

or more generally

$$P(B|\xi = a) = \frac{P(B \cap \xi = a)}{P(\xi = a)}$$

provided $P(\xi = a) > 0$. One of our goals is to seek a definition that makes sense when $P(\xi = a) = 0$. This involves dividing 0 by 0 and should involve differentiation of some kind. In the countable case we may think of $P(B|\xi = a_j)$ as a function $f_B(\xi)$ which is equal to $P(B|A_j)$ on $\xi = a_j$. We can rewrite our definition of

$$f_B(a_j) = P(B|\xi = a_j)$$

as

$$\int_{\xi = a_j} f_B(\xi)dP = P(B \cap \xi = a_j) \qquad \text{for each} \quad j$$

or summing over any arbitrary collection of j's

$$\int_{\xi\in E} f_B(\xi)dP = P(B\cap\{\xi\in E\})$$

Sets of the form $\xi \in E$ form a sub σ-field $\Sigma \subset \mathcal{F}$ and we can rewrite the definition as

$$\int_A f_B(\xi)dP = P(B\cap A)$$

for all $A \in \Sigma$. Of course in this case $A \in \Sigma$ if and only if A is a union of the atoms $\xi = a$ of the partition over a finite or countable subcollection of the possible values of a. Similar considerations apply to the conditional expectation of a random variable G given ξ. The equation becomes

$$\int_A g(\xi)dP = \int_A G(\omega)dP$$

or we can rewrite this as

$$\int_A g(\omega)dP = \int_A G(\omega)dP$$

for all $A \in \Sigma$ and instead of demanding that g be a function of ξ we demand that g be Σ measurable which is the same thing. Now the random variable ξ is out of the picture and rightly so. What is important is the information we have if we know ξ and that is the same if we replace ξ by a one-to-one function of itself. The σ-field Σ abstracts that information nicely. So it turns out that the proper notion of conditioning involves a sub σ-field $\Sigma \subset \mathcal{F}$. If G is an integrable function and $\Sigma \subset \mathcal{F}$ is given we will seek another integrable function g that is Σ measurable and satisfies

$$\int_A g(\omega)dP = \int_A G(\omega)dP$$

for all $A \in \Sigma$. We will prove existence and uniqueness of such a g and call it the conditional expectaion of G given Σ and denote it by $g = E[G|\Sigma]$.

The way to prove the above result will take us on a detour. A signed measure on a measurable space $(\Omega, \mathcal{F})$ is a set function $\lambda(.)$ defined for $A \in \mathcal{F}$ which is countably additive but not necessarily nonnegative. Countable addivity is again in any of the folowing two equivalent senses.

$$\lambda(\cup A_n) = \sum \lambda(A_n)$$

for any countable collection of disjoint sets in $\mathcal{F}$, or

$$\lim_{n\to\infty} \lambda(A_n) = \lambda(A)$$

whenver $A_n \downarrow A$ or $A_n \uparrow A$.

Examples of such λ can be constructed by taking the difference $\mu_1 - \mu_2$ of two nonnegative measures μ_1 and μ_2.

Definition 4.1. A set $A \in \mathcal{F}$ is totally positive (totally negative) for λ if for every subset $B \in \mathcal{F}$ with $B \subset A$ $\lambda(B) \geq 0$. (≤ 0)

Remark: A measurable subset of a totally positive set is totally positive. Any countable union of totally positive subsets is again totally positive.

Lemma 4.7. If λ is a countably additive signed measure on $(\Omega, \mathcal{F})$,

$$\sup_{A\in\mathcal{F}} |\lambda(A)| < \infty$$

Proof: The key idea in the proof is that, since $\lambda(\Omega)$ is a finite number, if $\lambda(A)$ is large so is $\lambda(A^c)$ with an opposite sign. In fact, it is not hard to see that $||\lambda(A)| - |\lambda(A^c)|| \leq |\lambda(\Omega)|$ for all $A \in \mathcal{F}$. Another fact is that if $\sup_{B\subset A} |\lambda(B)|$ and $\sup_{B\subset A^c} |\lambda(B)|$ are finite, so is $\sup_B |\lambda(B)|$. Now let us complete the proof. Given a subset $A \in \mathcal{F}$ with $\sup_{B\subset A} |\lambda(B)| = \infty$, and any positive number N, there is a subset $A_1 \in \mathcal{F}$ with $A_1 \subset A$ such that $|\lambda(A_1)| \geq N$ and $\sup_{B\subset A_1} |\lambda(A)| = \infty$. This is obvious because if we pick a set $E \subset A$ with $|\lambda(E)|$ very large so will $\lambda(E^c)$ be. At least one of the two sets E, E^c will have the second property and we can call it A_1. If we proceed by induction we have a sequence A_n that is $\downarrow$ and $|\lambda(A_n)| \to \infty$ that contradicts countable additivity.

Lemma 4.8. Given a subset $A \in \mathcal{F}$ with $\lambda(A) = \ell > 0$ there is a subset $\bar{A} \subset A$ that is totally positive with $\lambda(\bar{A}) \geq \ell$.

Proof: Let us define $m = \inf_{B\subset A} \lambda(B)$. Since the empty set is included, $m \leq 0$. If $m = 0$ then A is totally positive and and we are done. So let us assume that $m < 0$. By the previous lemma $m > -\infty$. Let us find $B_1 \subset A$ such that $\lambda(B_1) \leq \frac{m}{2}$. Then for $A_1 = A - B_1$ we have $A_1 \subset A$, $\lambda(A_1) \geq \ell$ and $\inf_{B\subset A_1} \lambda(B) \geq \frac{m}{2}$. By induction we can find A_k with

$A \supset A_1 \supset \cdots \supset A_k \cdots$, $\lambda(A_k) \geq \ell$ for every k and $\inf_{B \subset A_k} \lambda(A_k) \geq \frac{m}{2^k}$. Clearly if we define $\bar{A} = \cap A_k$ which is the decreasing limit, $\bar{A}$ works.

Theorem 4.9.(Hahn-Jordan Decomposition). Given a countably additive signed measure λ on $(\Omega, \mathcal{F})$ it can be written always as $\lambda = \mu^+ - \mu^-$ the difference of two nonnegative measures. Moreover μ^+ and μ^- may be chosen to be orthogonal i.e, there are disjoint sets $\Omega_+, \Omega_- \in \mathcal{F}$ such that $\mu^+(\Omega_-) = \mu^-(\Omega_+) = 0$. In fact Ω_+ and Ω_- can be taken to be subsets of Ω that are respectively totally positive and totally negative for λ. $\mu^\pm$ then become just the restrictions of λ to $\Omega_\pm$.

Proof: Totally positive sets are closed under countable unions, disjoint or not. Let us define $m^+ = \sup_A \lambda(A)$. If $m^+ = 0$ then $\lambda(A) \leq 0$ for all A and we can take $\Omega_+ = \Phi$ and $\Omega_- = \Omega$ which works. Assume that $m^+ > 0$. There exist sets A_n with $\lambda(A) \geq m^+ - \frac{1}{n}$ and therefore totally positive subsets $\bar{A}_n$ of A_n with $\lambda(\bar{A}_n) \geq m^+ - \frac{1}{n}$. Clearly $\Omega_+ = \cup_n \bar{A}_n$ is totally positive and $\lambda(\Omega_+) = m^+$. It is easy to see that $\Omega_- = \Omega - \Omega_+$ is totally negative. $\mu^\pm$ can be taken to be the restriction of λ to $\Omega_\pm$.

Remark: If $\lambda = \mu^+ - \mu^-$ with μ^+ and μ^- orthogonal to each other, then they have to be the restrictions of λ to the totally positive and totally negative sets for λ and such a representation for λ is unique. It is clear that in general the representation is not unique because we can add a common μ to both μ^+ and μ^- and the μ will cancel when we compute $\lambda = \mu^+ - \mu^-$.

Remark: If μ is a nonnegative measure and we define λ by

$$\lambda(A) = \int_A f(\omega)\, d\mu = \int \chi_A(\omega) f(\omega)\, d\mu$$

where f is an integrable function, then λ is a countably additive signed measure and $\Omega_+ = \{\omega : f(\omega) > 0\}$ and $\Omega_- = \{\omega : f(\omega) < 0\}$. If we define $f^\pm(\omega)$ as the positive and negative parts of f, then

$$\mu^\pm(A) = \int_A f^\pm(\omega)\, d\mu$$

The signed measure λ that was constructed in the preceeding remark enjoys a very special relationship to μ. For any set A with $\mu(A) = 0$,

$\lambda(A) = 0$ because the integrand $\chi_A(\omega)f(\omega)$ is 0 for μ-almost all ω and for all practical purposes is a function that vanishes identically.

Definition 4.2. A signed measure λ is said to be absolutely continuous with respect to a nonnegative measure μ, $\lambda << \mu$ in symbols, if whenever $\mu(A)$ is zero for a set $A \in \mathcal{F}$ it is also true that $\lambda(A) = 0$.

Theorem 4.10. **(Radon-Nikodym).** If $\lambda << \mu$ then there is an integrable function $f(\omega)$ such that

$$\lambda(A) = \int_A f(\omega)\, d\mu \tag{4.1}$$

for all $A \in \mathcal{F}$. The function f is uniquely determined almost everywhere and is called the Radon-Nikodym derivative of λ with respect to μ. It is denoted by

$$f(\omega) = \frac{d\lambda}{d\mu}$$

Proof: The proof depends on the decomposition theorem. We saw that if the relation 4.1 holds, then $\Omega_+ = \{\omega : f(\omega) > 0\}$. If we define $\lambda_a = \lambda - a\mu$, then λ_a is a signed measure for every real number a. Let us define $\Omega(a)$ to be the totally positive subset of λ_a. These sets are only defined up to sets of measure zero, and we can only handle a countable number of sets of measure 0 at one time. So it is prudent to restrict a to the set Q of rational numbers. Roughly speaking $\Omega(a)$ will be the sets $f(\omega) > a$ and we will try to construct f from the sets $\Omega(a)$ by the definition

$$f(\omega) = [\sup\ a \in Q : \omega \in \Omega(a)]$$

The plan is to check that the function $f(\omega)$ defined above works. Since λ_a is getting more negative as a increases, $\Omega(a)$ is $\downarrow$ as $a \uparrow$. There is trouble with sets of measure 0 for every comparison between two rationals a_1 and a_2. Collect all such troublesome sets (only a countable number and throw them away). In other words we may assume without loss of generality that $\Omega(a_1) \subset \Omega(a_2)$ whenever $a_1 > a_2$. Clearly

$$\begin{aligned}\{\omega : f(\omega) > x\} &= \{\omega : \omega \in \Omega(y) \text{ for some rational } y > x\}\\ &= \cup_{\substack{y>x\\ y\in Q}} \Omega(y)\end{aligned}$$

and this makes f measurable. If $A \subset \cap_a \Omega(a)$ then $\lambda(A) - a\mu(A) \geq 0$ for all A. If $\mu(A) > 0$, $\lambda(A)$ has to be infinite which is not possible. Therefore $\mu(A)$ has to be zero and by absolute continuity $\lambda(A) = 0$ as well. On the other hand if $A \cap \Omega(a) = \Phi$ for all a, then $\lambda(A) - a\mu(A) \leq 0$ for all a and again if $\mu(A) > 0$, $\lambda(A) = -\infty$ which is not possible either. Therefore $\mu(A)$, and by absolute continuity, $\lambda(A)$ are zero. This proves that $f(\omega)$ is finite almost everywhere with rsepect to both λ and μ. Let us take two real numbers $a < b$ and consider $E_{a,b} = \{\omega : a \leq f(\omega) \leq b\}$. It is clear that the set $E_{a,b}$ is in $\Omega(a')$ and $\Omega^c(b')$ for any $a' < a$ and $b' > b$. Therefore for any set $A \subset E_{a,b}$ by letting a' and b' tend to a and b

$$a\,\mu(A) \leq \lambda(A) \leq b\,\mu(A)$$

Now we are essentially done. Let us take a grid $\{n\,h\}$ and consider $E_n = \{\omega : nh \leq f(\omega) < (n+1)h\}$ for $-\infty < n < \infty$. Then for any $A \in \mathcal{F}$ and each n,

$$\begin{aligned}\lambda(A \cap E_n) - h\mu(A \cap E_n) &\leq n\,h\,\mu(A \cap E_n) \leq \int_{A \cap E_n} f(\omega)\,d\mu \\ &\leq (n+1)\,h\mu(A \cap E_n) \leq \lambda(A \cap E_n) + h\,\mu(A \cap E_n)\end{aligned}$$

Summing over n we have

$$\lambda(A) - h\,\mu(A) \leq \int_A f(\omega)\,d\mu \leq \lambda(A) + h\,\mu(A)$$

proving the integrability of f and if we let $h \to 0$ establishing

$$\lambda(A) = \int_A f(\omega)\,d\mu$$

for all $A \in \mathcal{F}$.

Remark: (Uniqueness). If we have two choices of f say f_1 and f_2 their difference $g = f_1 - f_2$ satisfies

$$\int_A g(\omega)\,d\mu = 0$$

for all $A \in \mathcal{F}$. If we take $A_\epsilon = \{\omega : g(\omega) \geq \epsilon\}$, then $0 \geq \epsilon\mu(A_\epsilon)$ and this implies $\mu(A_\epsilon) = 0$ for all $\epsilon > 0$ or $g(\omega) \leq 0$, almost everywhere with respect to μ. A similar argument establishes $g(\omega) \geq 0$ almost everywhere with respect to μ. Therefore $g = 0$ a.e. μ proving uniqueness.

Exercise 4.1. If $\lambda(A) \geq 0$ for all $A \in \mathcal{F}$, prove that $f(\omega) \geq 0$ almost everywhere.

Exercise 4.2. If Ω is a countable set and $\mu(\{\omega\}) > 0$ for each single point set prove that any measure λ is absolutely continuous with respect to λ and calculate the Radon-Nikodym derivative.

Exercise 4.3. Let $F(x)$ be a distribution function on the line with $F(0) = 0$ and $F(1) = 1$ so that the probability measure α corresponding to it lives on the interval $[0,1]$. If $F(x)$ satisfies a Lipschitz condition

$$|F(x) - F(y)| \leq A|x - y|$$

then prove that $\alpha << m$ where m is the Lebesgue measure on $[0,1]$. Show also that $0 \leq \frac{d\alpha}{dm} \leq A$ almost surely.

Exercise 4.4. If ν, λ, μ are three nonnegative measures such that $\nu << \lambda$ and $\lambda << \mu$ then show that $\nu << \mu$ and

$$\frac{d\nu}{d\mu} = \frac{d\nu}{d\lambda}\frac{d\lambda}{d\mu}$$

a.e.

Exercise 4.5. If λ, μ are nonnegative measures with $\lambda << \mu$ and $\frac{d\lambda}{d\mu} = f$, then show that g is integrable with respect to λ if and only if $g\,f$ is integrable with respect to μ and

$$\int g(\omega)\,d\lambda = \int g(\omega)\,f(\omega)\,d\mu$$

Exercise 4.6. Given two nonnegative measures λ and μ, λ is said to be uniformly absolutely continuous with respect to μ on $\mathcal{F}$ if for any $\epsilon > 0$ there exists a $\delta > 0$ such that for any $A \in \mathcal{F}$ with $\mu(A) < \delta$ it is true that $\lambda(A) < \epsilon$. Use the Radon-Nikodym theorem to show that absolute continuity on a σ-field $\mathcal{F}$ implies uniform absolute continuity. If $\mathcal{F}_0$ is a field that generates the σ-field $\mathcal{F}$ show by an example that absolute continuity on $\mathcal{F}_0$ does not imply absolute continuity on $\mathcal{F}$. Show however that uniform absolute continuity on $\mathcal{F}_0$ implies uniform absolute continuity and therefore absolute continuity on $\mathcal{F}$.

Exercise 4.7. If F is a distribution function on the line show that it is absolutely continuous with respect to Lebesgue measure on the line, if and only if for any $\epsilon > 0$, there exists a $\delta > 0$ such that for arbitrary finite collection of disjoint intervals $I_j = [a_j, b_j]$ with $\sum_j |b_j - a_j| < \delta$ it follows that $\sum_j [F(b_j) - F(a_j)] \leq \epsilon$.

Conditional Probability

In the Radon-Nikodym theorem, if $\lambda << \mu$ are two probability distributions on $(\Omega, \mathcal{F})$, we defined the Radon-Nikodym derivative $f(\omega) = \frac{d\lambda}{d\mu}$ as an $\mathcal{F}$ measurable function such that

$$\lambda(A) = \int_A f(\omega)\, d\mu \qquad \text{for all} \quad A \in \mathcal{F}$$

If $\Sigma \subset \mathcal{F}$ is a sub σ-field, the absolute continuity of λ with respect to μ on Σ is clearly implied by the absolute continuity of λ with respect to μ on $\mathcal{F}$. We can therefore apply the Radon-Nikodym theorem on the measurable space (Ω, Σ), and we will obtain a new Radon-Nikodym derivative

$$g(\omega) = \frac{d\lambda}{d\mu} = \frac{d\lambda}{d\mu}\Big|_{\Sigma}$$

such that

$$\lambda(A) = \int_A g(\omega)\, d\mu \qquad \text{for all} \quad A \in \Sigma$$

and g is Σ measurable. Since the old function $f(\omega)$ was only $\mathcal{F}$ measurable, in general, it cannot be used as the Radon-Nikodym derivative for the sub σ-field Σ. Now if f is an integrable function on $(\Omega, \mathcal{F}, \mu)$ and $\Sigma \subset \mathcal{F}$ is a sub σ-field we can define λ on $\mathcal{F}$ by

$$\lambda(A) = \int_A f(\omega)\, d\mu \qquad \text{for all} \quad A \in \mathcal{F}$$

and recalculate the Radon-Nikodym derivative g for Σ and g will be a Σ measurable, integrable function such that

$$\lambda(A) = \int_A g(\omega)\, d\mu \qquad \text{for all} \quad A \in \Sigma$$

In other words g is the perfect candidate for the conditional expectation

$$g(\omega) = E\{f(\cdot)|\Sigma\}$$

We have therefore proved the existence of the conditional expectation.
Theorem 4.11. The conditional expectation as mapping of $f \to g$ has the following properties.
(i). If $g = E\{f|\Sigma\}$ then $E[g] = E[f]$. $E[1|\Sigma] = 1$ a.e.
(ii). If f is nonnegative then $g = E\{f|\Sigma\}$ is almost surely nonnegative.
(iii). The map is linear. If a_1, a_2 are constants

$$E\{a_1 f_1 + a_2 f_2|\Sigma\} = a_1 E\{f_1|\Sigma\} + a_2 E\{f_2|\Sigma\} \quad \text{a.e.}$$

(iv). If $g = E\{f|\Sigma\}$, then

$$\int |g(\omega)|\, d\mu \leq \int |f(\omega)|\, d\mu$$

(v). If h is a bounded Σ measurable function then

$$E\{f\, h|\Sigma\} = h\, E\{f|\Sigma\} \quad \text{a.e.}$$

(vi). If $\Sigma_2 \subset \Sigma_1 \subset \mathcal{F}$, then

$$E\{f|\Sigma_2\} = E\{E\{f|\Sigma_1\}|\Sigma_2\}$$

(vii). If $\phi(x)$ is a convex function of x, and $g = E\{f|\Sigma\}$ then

$$E\{\phi(f(\omega))|\Sigma\} \geq \phi(g(\omega)) \quad \text{a.e.}$$

and if we take expectations

$$E[\phi(f)] \geq E[\phi(g)].$$

Proof: (i), (ii) and (iii) are obvious. For (iv) we note that if $d\lambda = f\, d\mu$

$$\int |f|\, d\mu = \sup_{A \in \mathcal{F}} \lambda(A) - \inf_{A \in \mathcal{F}} \lambda(A)$$

and if we replace $\mathcal{F}$ by a sub σ-field Σ the right hand side is decreased. (v) is ovious if h is the indicator function of a set A in Σ. To go from indicator functions to simple functions to bounded measurable functions is routine. (vi) is an easy consequence of the definition. To see (vii) we note that if $f_1 \geq f_2$ then $E\{f_1|\Sigma\} \geq E\{f_2|\Sigma\}$ a.e. and consequently $E\{\max(f_1, f_2)|\Sigma\} \geq \max(g_1, g_2)$ a.e. where $g_i = E\{f_i|\Sigma\}$ for $i = 1, 2$. Since we can represent any convex function ϕ as $\phi(x) = \sup_a [ax - \psi(a)]$, limiting ourselves to rational a, we have only a countable set of functions to deal with, and

$$\begin{aligned} E\{\phi(f)|\Sigma\} &= E\{\sup_a [af - \psi(a)]|\Sigma\} \\ &\geq \sup_a [a\, E\{f|\Sigma\} - \psi(a)] \\ &= \sup_a [a\, g - \psi(a)] \\ &= \phi(g) \end{aligned}$$

a.e. and after taking expectations

$$E[\phi(f)] \geq E[\phi(g)]$$

In a somewhat more familiar context if $\mu = \lambda_1 \times \lambda_2$ is a product measure on $(\Omega, \mathcal{F}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ and we take $\Sigma = \{A \times \Omega_2 : A \in \mathcal{F}_1\}$ then for any function $f(\omega) = f(\omega_1, \omega_2)$, $E[f(\cdot)|\Sigma] = g(\omega)$ where $g(\omega) = g(\omega_1)$ is given by

$$g(\omega_1) = \int_{\Omega_2} f(\omega_1, \omega_2)\, d\lambda_2$$

so that the conditional expectation is just integrating the unwanted variable ω_2. We can go one step more. If $\phi(x, y)$ is the joint density on R^2 of two random variables X, Y (with respect to the Lebesgue measure on R^2), and $\psi(x)$ is the marginal density of X given by

$$\psi(x) = \int_{-\infty}^{\infty} \phi(x, y)\, dy$$

then for any integrable function $f(x, y)$

$$E[f(X,Y)|X] = E[f(\cdot,\cdot)|\Sigma] = \frac{\int_{-\infty}^{\infty} f(x,y)\, \phi(x,y)\, dy}{\psi(x)}$$

where Σ is the σ-field of vertical strips $A \times (-\infty, \infty)$ with a measurable horizontal base A.

Exercise 4.8. If f is already Σ measurable then $E[f|\Sigma] = f$. This suggests that the map $f \to g = E[f|\Sigma]$ is some sort of a projection. In fact if we consider the Hilbert space $\mathbf{H} = L_2[\Omega, \mathcal{F}, \mu]$ of all $\mathcal{F}$ measurable square integrable functions with an inner product

$$< f, g >_\mu = \int f\, g\, d\mu$$

then

$$\mathbf{H}_0 = L_2[\Omega, \Sigma, \mu] \subset \mathbf{H} = L_2[\Omega, \mathcal{F}, \mu]$$

and $f \to E[f|\Sigma]$ is seen to be the same as the orthogonal projection from $\mathbf{H}$ onto $\mathbf{H}_0$. Prove it.

Conditional expectation is then the best nonlinear predictor if the loss function is the expected (mean) square error.

We now turn our attention to conditional probability. If we take $f = \chi_B(\omega)$ then $E[f|\Sigma] = P(\omega, B)$ is called the conditional probability of B given Σ. It is characterized by the property that it is Σ measurable as a function of ω and for any $A \in \Sigma$

$$\mu(A \cap B) = \int_A P(\omega, B)\, d\mu$$

Theorem 4.12. *$P(\cdot, \cdot)$ has the following properties.*
(i). $P(\omega, \Omega) = 1, P(\omega, \Phi) = 0$ a.e.
(ii). For any $B \in \mathcal{F}$, $0 \le P(\omega, B) \le 1$ a.e.
(iii). For any countable collection $\{B_j\}$ of disjoint sets in $\mathcal{F}$,

$$P(\omega, \cup_j B_j) = \sum_j P(\omega, B_j) \quad \text{a.e.}$$

(iv). If $B \in \Sigma$, $P(\omega, B) = \chi_B(\omega)$ a.e.

Proof: All are easy consequences of properties of conditional expectations. Property (iii) perhaps needs an explanation. If $E[|f_n - f|] \to 0$ by the properties of conditional expectation $E[|E\{f_n|\Sigma\} - E\{f|\Sigma\}| \to 0$. Property (iii) is an easy consequence of this.

The problem with the above theorem is that every property is valid only almost everywhere. There are exceptional sets of measure zero for each case. While each null set or a countable number of them can be ignored we have an uncountable number of null sets and we would like a single null set outside which all the properties hold. This means constructing a good version of the conditional probability. It may not be always possible. If possible, such a version is called a regular conditional probability. The existence of such a regular version depends on the space $(\Omega, \mathcal{F})$ and the sub σ-field Σ being nice. If Ω is a complete separable metric space and $\mathcal{F}$ are its Borel stes, and if Σ is any countably generated sub σ-field of $\mathcal{F}$, then it is nice enough. We will prove it in the special case when $\Omega = [0,1]$ is the unit interval and $\mathcal{F}$ are the Borel subsets of $[0,1]$. Σ can be any countably generated sub σ-field of $\mathcal{F}$.

The trick is not to be too ambitious in the first place and try to construct the conditional expectations

$$Q(\omega, B) = E\{\chi_B(\omega)|\Sigma\}$$

only for sets B given by $B = (-\infty, x)$ for rational x. We call our conditional expectation, which is in fact a conditional probability, by $F(\omega, x)$. By the properties of conditional expectations for any pair of rationals $x < y$, there is a null set $E_{x,y}$, such that for $\omega \notin E_{x,y}$

$$F(\omega, x) \le F(\omega, y)$$

Moreover for any rational $x < 0$, there is a null set N_x outside which $F(\omega, x) = 0$ and similar null sets N_x for $x > 1$, ouside which $F(\omega, x) = 1$. If we collect all these null sets, of which there are only countably many, and take their union, we get a null set $N \in \Sigma$ such that for $\omega \notin N$, we have have a family $F(\omega, x)$ defined for rational x that satisfies

$$F(\omega, x) \le F(\omega, y) \quad \text{if} \quad x < y \quad \text{are rational}$$

$$\begin{aligned} F(\omega, x) &= 0 \quad \text{for rational} \quad x < 0 \quad \text{and} \quad F(\omega, x) \\ &= 1 \quad \text{for rational} \quad x > 1 \end{aligned}$$

$$P(A \cap [0,x]) = \int_A F(\omega, x)\, dP \quad \text{for all} \quad A \in \Sigma$$

For $\omega \notin N$ and real y we can define

$$G(\omega, y) = \lim_{\substack{x \downarrow y \\ x \text{ rational}}} F(\omega, x)$$

For $\omega \notin N$, G is a right continuous nondecreasing function (distribution function) with $G(y) = 0$ for $y < 0$ and $G(y) = 1$ for $y \geq 1$. There is then a probability measure $\hat{Q}(\omega, B)$ on the Borel subsets of $[0,1]$ such that $\hat{Q}(\omega, [0, y]) = G(\omega, y)$ for all y. $\hat{Q}$ is our candidate for regular conditional probability. Clearly $\hat{Q}(\omega, I)$ is Σ measurable for all intervals I and by standard arguments will continue to be Σ measurable for all Borel sets $B \in \mathcal{F}$. If we check that

$$P(A \cap [0, x]) = \int_A G(\omega, x)\, dP \quad \text{for all} \quad A \in \Sigma$$

for all $0 \leq x \leq 1$ then

$$P(A \cap I) = \int_A \hat{Q}(\omega, I)\, dP \quad \text{for all} \quad A \in \Sigma$$

for all intervals I and by standard arguments this will extend to finite disjoint unions of half open intervals that constitute a field and finally to the σ-field $\mathcal{F}$ generated by that field. To verify that for all real y,

$$P(A \cap [0, y]) = \int_A G(\omega, y)\, dP \quad \text{for all} \quad A \in \Sigma$$

we start from

$$P(A \cap [0, x]) = \int_A F(\omega, x)\, dP \quad \text{for all} \quad A \in \Sigma$$

valid for rational x and let $x \downarrow y$ through rationals. From the countable additivity of P the left hand side converges to $P(A \cap [0, y])$ and by the bounded convergence theorem, the right hand side converges to $\int_A G(\omega, y)\, dP$ and we are done. Finally from the uniqueness of the conditional expectation if $A \in \Sigma$

$$\hat{Q}(\omega, A) = \chi_A(\omega)$$

provided $\omega \notin N_A$, which is a null set that depends on A. We can take a countable set Σ_0 of generators A that forms a field and get a single null set N such that if $\omega \notin N$

$$\hat{Q}(\omega, A) = \chi_A(\omega)$$

for all $A \in \Sigma_0$. Since both side are countably additive measures in A and as they agree on Σ_0 they have to agree on Σ as well.

Exercise 4.9. **Disintegration Theorem.** Let μ be a probability measure on the plane R^2 with a marginal distribution α for the first coordinate. In other words if we denote by (x, y) a point in R^2, then

$$\int_{R^2} f(x)\, d\mu = \int_R f(x)\, d\alpha$$

Show that there exist measures β_x depending measurably on x such that $\beta_x[\{x\} \times R] = 1$ i.e. β_x is supported on the vertial line through $(x, y) : y \in R$ and $\mu = \int_R \beta_x\, d\alpha$. The converse is of course easier. Given α and β_x we can construct a unique μ such that μ disintegrates as expected.